

COMMUNICATIONS

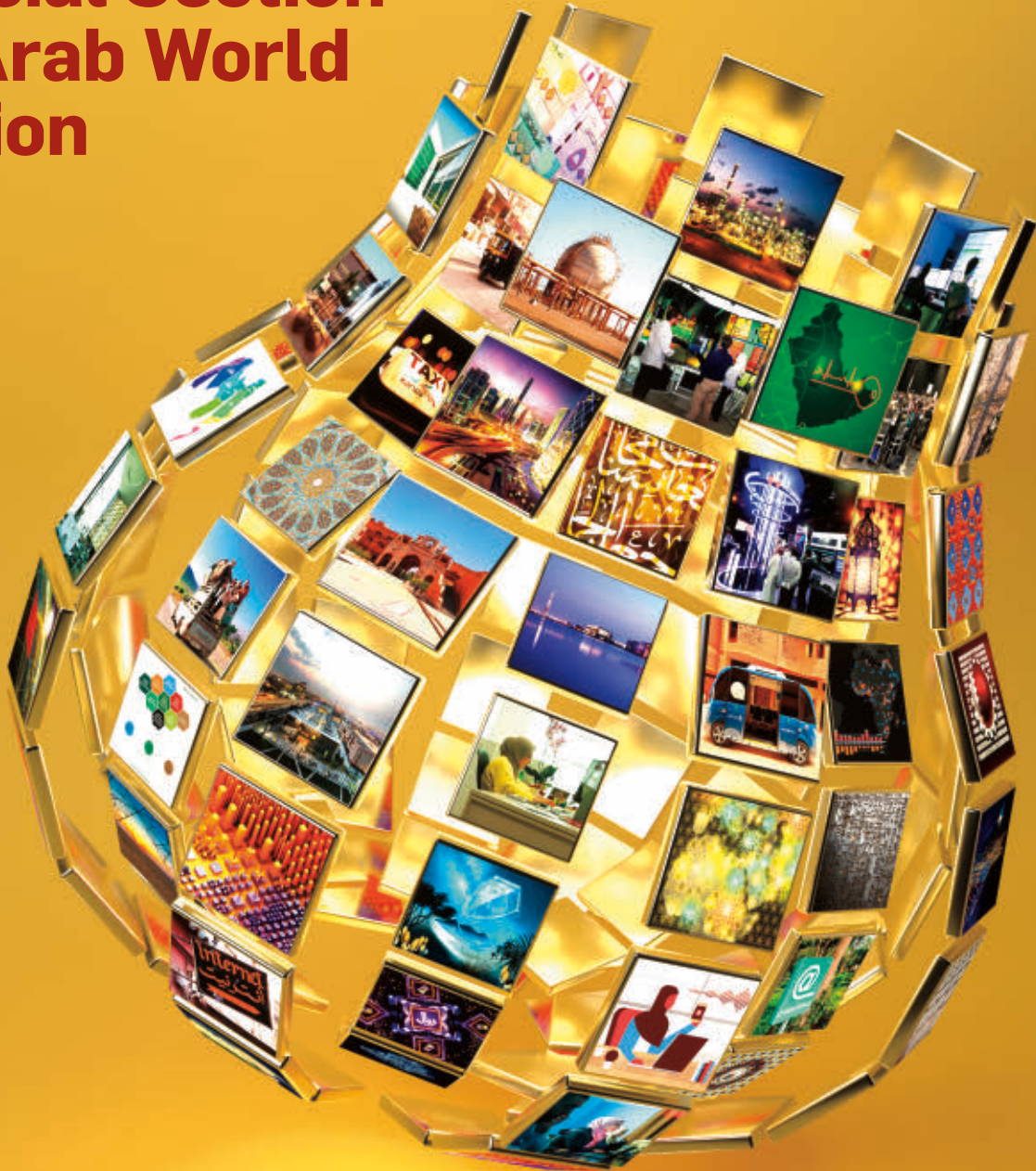
CACM.ACM.ORG

OF THE

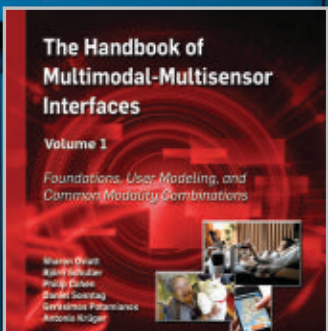
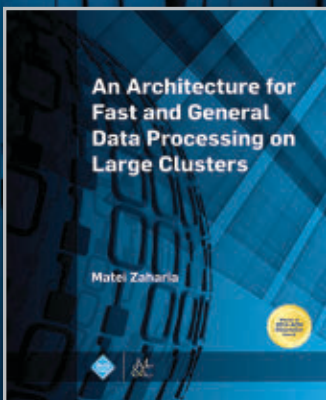
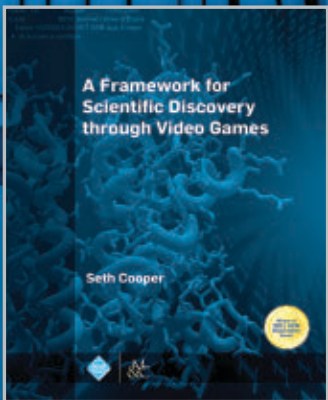
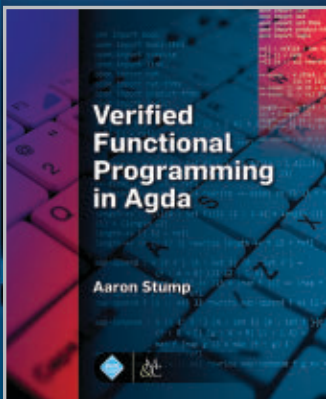
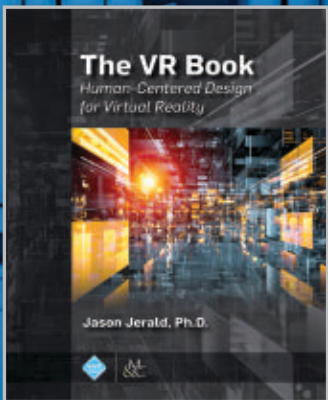
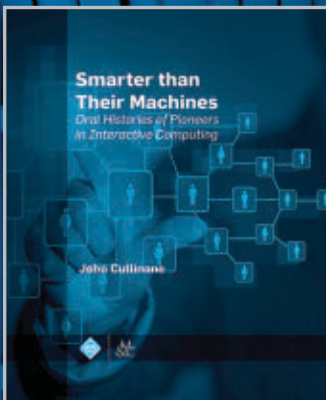
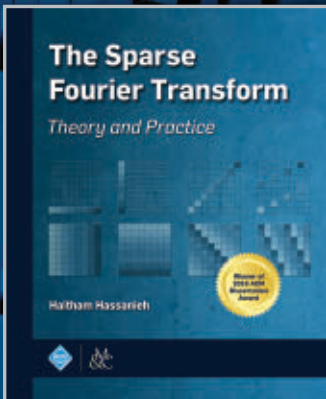
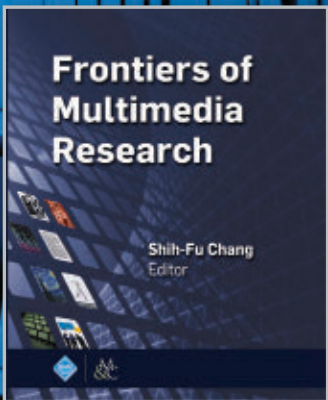
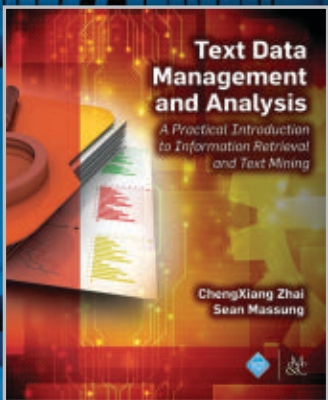
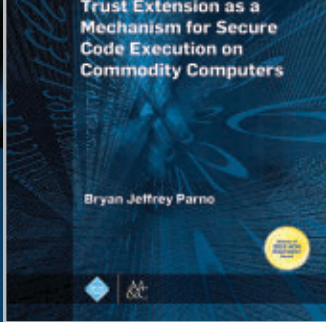
ACM

04/2021 VOL.64 NO.04

Special Section on Arab World Region



Succinct Range Filters
Safe Systems Programming in Rust
The Worsening State of Ransomware
When Hackers Were Heroes



In-depth. Innovative. Insightful.

Inspired by the need for high-quality computer science publishing at the graduate, faculty, and professional levels, ACM Books are affordable, current, and comprehensive in scope.

**Full Collection | Title List
Now Available**

For more information, please visit
<http://books.acm.org>



Association for Computing Machinery
1601 Broadway, 10th Floor, New York, NY 10019-7434, USA
Phone: +1-212-626-0658 Email: acmbooks-info@acm.org



*Making Waves,
Combining Strengths*

CHI 2021
May 8-13, 2021



In 2021 we are offering reduced registration rates for everyone and even further reduced rates for individuals residing in 139 economically developing countries who take advantage of early registration.

You can find more details on chi2021.acm.org

General Co-Chair
Yoshifumi Kitamura
Aaron Quigley

Technical Program Chairs
Katherine Isbister
Takeo Igarashi

Papers Chairs
Pernille Bjorn
Steven Drucker

Deadlines

March 16th : Early Registration Ends

Apr 13th : Standard Registration Ends

Late Registration : Up to CHI 2021

**Online Virtual Conference
(originally Yokohama, Japan)**

Departments

- 5 **Cerf's Up**
What Does a Static, Sustainable Economy Look Like?
By Vinton G. Cerf
-
- 6 **BLOG@CACM**
The SolarWinds Hack, and a Grand Challenge for CS Education
John Arquilla analyzes the latest in a long line of cyber intrusions, while Mark Guzdial considers how to prepare CS students for professional decision making.

174 **Careers**

Last Byte

- 176 **Upstart Puzzles**
Roulette Angel
Where will the ball drop when the spinning roulette wheel stops?
By Dennis Shasha

News



- 9 **The Best of NLP**
Natural language processing delves more deeply into its knowledge gap.
By Chris Edwards
-
- 12 **Deep Learning Speeds MRI Scans**
Machine intelligence significantly reduces the time needed for an MRI scan, which can help reduce patient anxiety.
By Paul Marks
-
- 15 **The Worsening State of Ransomware**
Sophisticated, dangerous ransomware is the new normal ... and there is no simple fix.
By Samuel Greengard

Viewpoints

- 20 **Technology Strategy and Management From Remote Work to Working From Anywhere**
Tracing temporary work modifications resulting in permanent organizational changes.
By Mari Sako
-
- 23 **Broadening Participation**
Reflections on Black in Computing
Seeking to improve systemic fairness in the computing realm.
By Quincy Brown, Tyrone Grandison, Jamika D. Burge, Odest Chadwicke Jenkins, and Tawanna Dillahunt
-
- 25 **Kode Vicious**
The Non-Psychopath's Guide to Managing an Open Source Project
Respect your staff, learn from others, and know when to let go.
By George V. Neville-Neil
-
- 28 **Historical Reflections**
When Hackers Were Heroes
The complex legacy of Steven Levy's obsessive programmers.
By Thomas Haigh
-
- 35 **Viewpoint**
Roots of 'Program' Revisited
Considering the fundamental nature and malleability of programming.
By Liesbeth De Mol and Maarten Bullynck
-
- 38 **Viewpoint**
Building a Multilingual Wikipedia
Seeking to develop a multilingual Wikipedia where content can be shared among language editions.
By Denny Vrandečić

Special Section



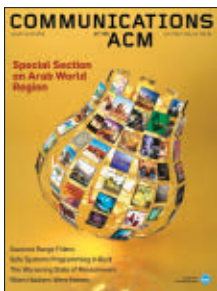
59

42 Arab World Region

This issue features an all-encompassing look at the latest technology advancements, achievements, and challenges emanating from the countries that make up Arab world.



Watch the co-organizers discuss this section in the exclusive *Communications* video. <https://cacm.acm.org/videos/arab-world-region>



About the Cover: This collection of articles is the latest in *Communications*'s series of regional special sections that began with the November 2018 issue and it represents a complete global go-round spotlighting news and technologies by region. Cover illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: KAUST photos courtesy of King Abdullah University of Science and Technology, OCRI photos courtesy of Qatar Computing Research Institute, Summit photo by Mahmoud Khales/Vinhua, courtesy of www.vinhua.com, CHI19 photo courtesy of ArabHCL.org, ArabCHI poster courtesy of ArabHCL.org, Rickshaw photo by Evranovostro/Shutterstock.com; Truck photo by Natalia Davidovich/Shutterstock.com; Taxi photo by Philip Lange/Shutterstock.com; Shuttle photo by Anna Ostanina/Shutterstock.com; Mosaic photo by fkaymak/Shutterstock.com; Doha photo by HasanZaidi/Shutterstock.com; AUC photo by Kazzaan/Shutterstock.com; Class photo by Momen_frames/Shutterstock.com. Additional stock images from Shutterstock.com.

Practice



130

130 Everything VPN Is New Again

The 24-year-old security model has found a second wind.
By David Crawshaw



Articles' development led by [acmqueue.queue.acm.org](https://queue.acm.org)

Contributed Articles

136 The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making

What does it mean to be fair?
By Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian

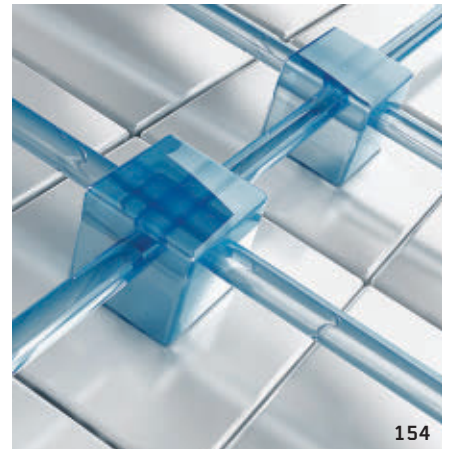
144 Safe Systems Programming in Rust

The promise and the challenges of the first industry-supported language to master the trade-off between safety and control.
By Ralf Jung, Jacques-Henri Jourdan, Robbert Krebbers, and Derek Dreyer



Watch the authors discuss this article in the exclusive *Communications* video. <https://cacm.acm.org/videos/programming-in-rust>

Review Articles



154

154 Transformers Aftermath: Current Research and Rising Trends

Attention, particularly self-attention, is a standard in current NLP literature, but to achieve meaningful models, attention is not enough.
By Eduardo Souza Dos Reis, Cristiano André Da Costa, Diórgenes Eugênio Da Silveira, Rodrigo Simon Bavaresco, Rodrigo Da Rosa Righi, Jorge Luis Victória Barbosa, Rodolfo Stoffel Antunes, Márcio Miguel Gomes, and Gustavo Federizzi

Research Highlights

165 Technical Perspective

The Strength of SuRF
By Stratos Idreos

166 Succinct Range Filters

By Huan Chen Zhang, Hyeontaek Lim, Viktor Leis, David G. Andersen, Michael Kaminsky, Kimberly Keeton, and Andrew Pavlo



ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

Executive Director and CEO

Vicki L. Hanson

Deputy Executive Director and COO

Patricia Ryan

Director, Office of Information Systems

Wayne Graves

Director, Office of Financial Services

Darren Ramdin

Director, Office of SIG Services

Donna Cappel

Director, Office of Publications

Scott E. Delman

ACM COUNCIL

President

Gabriele Kotsis

Vice-President

Joan Feigenbaum

Secretary/Treasurer

Elisa Bertino

Past President

Cherri M. Pancake

Chair, SGB Board

Jeff Jortner

Co-Chairs, Publications Board

Jack Davidson and Joseph Konstan

Members-at-Large

Nancy M. Amato; Tom Crick;

Susan Dumais; Mehran Sahami;

Alejandro Saucedo

SGB Council Representatives

Sarita Adve and Jeanna Neefe Matthews

BOARD CHAIRS

Education Board

Mehran Sahami and Jane Chu Prey

Practitioners Board

Terry Coatta

REGIONAL COUNCIL CHAIRS

ACM Europe Council

Chris Hankin

ACM India Council

Abhiram Ranade

ACM China Council

Wenguang Chen

PUBLICATIONS BOARD

Co-Chairs

Jack Davidson and Joseph Konstan

Board Members

Jonathan Aldrich; Phoebe Ayers;

Chris Hankin; Mike Heroux; James Larus;

Tulika Mitra; Marc Najork;

Michael L. Nelson; Theo Schlossnagle;

Eugene H. Spafford; Divesh Srivastava;

Bhavani Thuraisin; Robert Walker;

Julie R. Williamson

ACM U.S. Technology Policy Office

Adam Eisgrau

Director of Global Policy and Public Affairs

1701 Pennsylvania Ave NW, Suite 200,

Washington, DC 20006 USA

T (202) 580-6555; acmpo@acm.org

Computer Science Teachers Association

Jake Baskin

Executive Director

COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

Communications of the ACM is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

STAFF

DIRECTOR OF PUBLICATIONS

Scott E. Delman

cacm-publisher@cacm.acm.org

Executive Editor

Diane Crawford

Managing Editor

Thomas E. Lambert

Senior Editor

[vacant]

Senior Editor/News

Lawrence M. Fisher

Web Editor

David Roman

Editorial Assistant

Danbi Yu

Art Director

Andrij Borys

Associate Art Director

Margaret Gray

Assistant Art Director

Mia Angelica Balaquiot

Production Manager

Bernadette Shade

Intellectual Property Rights Coordinator

Barbara Ryan

Advertising Sales Account Manager

Ilia Rodriguez

Columnists

David Anderson; Michael Cusumano;

Peter J. Denning; Mark Guzdial;

Thomas Haigh; Leah Hoffmann; Mari Sako;

Pamela Samuelson; Marshall Van Alstyne

CONTACT POINTS

Copyright permission

permissions@hq.acm.org

Calendar items

calendar@cacm.acm.org

Change of address

acmhelp@acm.org

Letters to the Editor

letters@cacm.acm.org

WEBSITE

http://cacm.acm.org

WEB BOARD

Chair

James Landay

Board Members

Marti Hearst; Jason I. Hong;

Jeff Johnson; Wendy E. MacKay

AUTHOR GUIDELINES

http://cacm.acm.org/about-communications/author-center

ACM ADVERTISING DEPARTMENT

1601 Broadway, 10th Floor

New York, NY 10019-7434 USA

T (212) 626-0686

F (212) 869-0481

Advertising Sales Account Manager

Ilia Rodriguez

ilia.rodriguez@hq.acm.org

Media Kit acmm mediasales@acm.org

Association for Computing Machinery (ACM)

1601 Broadway, 10th Floor

New York, NY 10019-7434 USA

T (212) 869-7440; F (212) 869-0481

EDITORIAL BOARD

EDITOR-IN-CHIEF

Andrew A. Chien

aic@cacm.acm.org

Deputy to the Editor-in-Chief

Morgan Denlow

cacm.deputy.to.aic@gmail.com

SENIOR EDITOR

Moshe Y. Vardi

NEWS

Co-Chairs

Marc Snir and Alain Chesnais

Board Members

Tom Conte; Monica Divitini; Mei Kobayashi;

Rajeev Rastogi; François Sillion

VIEWPOINTS

Co-Chairs

Tim Finin; Susanne E. Hambrusch;

John Leslie King

Board Members

Virgilio Almeida; Terry Benzel; Michael L. Best;

Judith Bishop; Lorrie Cranor; Boi Falting;

James Grimmelmann; Mark Guzdial;

Haym B. Hirsch; Anupam Joshi; Richard Ladner;

Carl Landwehr; Beng Chin Ooi; Francesca Rossi;

Len Shustek; Loren Terveen; Marshall Van

Alstyne; Jeannette Wing; Susan J. Winter

PRACTICE

Co-Chairs

Stephen Bourne and Theo Schlossnagle

Board Members

Eric Allman; Samy Bahra; Peter Bailis;

Betsy Beyer; Terry Coatta; Stuart Feldman;

Nicole Forsgren; Camille Fournier;

Jessie Frazelle; Benjamin Fried; Tom Killalea;

Tom Limoncelli; Kate Matsudaira;

Marshall Kirk McKusick; Erik Meijer;

George Neville-Neil; Jim Waldo;

Meredith Whittaker

CONTRIBUTED ARTICLES

Co-Chairs

James Larus and Gail Murphy

Board Members

Robert Austin; Kim Bruce; Alan Bundy;

Peter Buneman; Premkumar T. Devanbu;

Jane Cleland-Huang; Yannis Ioannidis;

Trent Jaeger; Somesh Jha; Gal A. Kaminka;

Ben C. Lee; Igor Markov; m.c. schraefel;

Hannes Werthner; Reinhard Wilhelm;

Rich Wolski

RESEARCH HIGHLIGHTS

Co-Chairs

Shriram Krishnamurthi

and Orna Kupferman

Board Members

Martin Abadi; Amr El Abbadi;

Animashree Anandkumar; Sanjeev Arora;

Michael Backes; Maria-Florina Balcan;

Azer Bestavros; David Brooks; Stuart K. Card;

Jon Crowcroft; Lieven Eeckhout;

Alexei Efron; Bryan Ford; Alon Halevy;

Rennot Heiser; Takeo Igarashi;

Srinivasan Keshav; Sven Koenig;

Ran Libeskind-Hadas; Karen Liu;

Tim Roughgarden; Guy Steele, Jr.;

Robert Williamson; Margaret H. Wright;

Nicholai Zeldovich; Andreas Zeller

SPECIAL SECTIONS

Co-Chairs

Sriram Rajamani, Haibo Chen,

and P. J. Narayanan

Board Members

Sue Moon; Tao Xie; Kenjiro Taura; David Padua

ACM Copyright Notice

Copyright © 2021 by Association for Computing Machinery, Inc. (ACM). Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or fee. Request permission to publish from permissions@hq.acm.org or fax (212) 869-0481.

For other copying of articles that carry a code at the bottom of the first or last page or screen display, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center; www.copyright.com.

Subscriptions

An annual subscription cost is included in ACM member dues of \$99 (\$40 of which is allocated to a subscription to *Communications*); for students, cost is included in \$42 dues (\$20 of which is allocated to a *Communications* subscription). A nonmember annual subscription is \$269.

ACM Media Advertising Policy

Communications of the ACM and other ACM Media publications accept advertising in both print and electronic formats. All advertising in ACM Media publications is at the discretion of ACM and is intended to provide financial support for the various activities and services for ACM members. Current advertising rates can be found by visiting <http://www.acm-media.org> or by contacting ACM Media Sales at (212) 626-0686.

Single Copies

Single copies of *Communications of the ACM* are available for purchase. Please contact acmhelp@acm.org.

COMMUNICATIONS OF THE ACM

(ISSN 0001-0782) is published monthly by ACM Media, 1601 Broadway, 10th Floor New York, NY 10019-7434 USA. Periodicals postage paid at New York, NY 10001, and other mailing offices.

POSTMASTER

Please send address changes to *Communications of the ACM* 1601 Broadway, 10th Floor New York, NY 10019-7434 USA

Printed in the USA.



Association for Computing Machinery





Vinton G. Cerf

DOI:10.1145/3450610

What Does a Static, Sustainable Economy Look Like?

I AM NOT AN economist. Come to think of it, I am not a lawyer either. But I would like to speculate a bit about what a static, sustainable economy might look like—or, at least, what conditions might have to be satisfied for such an economy to be realistic. On the presumption our present economy is heavily rooted in single-use packaging, replace not repair, and growing populations to drive markets, we might conclude our consumption of non-renewable resources will be decreasingly sustainable. It is already apparent the most “advanced” economies are consuming resources far above their nominal share based on population. For a long time, a growing population assured increased consumption and thus an increasing GDP. I will set aside a rant about the inequality and inequity of income and wealth distribution as this is not germane to the more central question whether a static economy is sustainable or even desirable.

On the presumption the consumption of non-renewable resources is not sustainable, we might ask—as I have in previous columns—whether reparability and longevity of products and services might become increasingly important. Software may have a role to play here. The Tesla electric vehicle is a useful example. It increases its functionality and utility through new software-based capabilities. This extends the useful life of the vehicle.

Design for reparability and circular recycling strike me as elements of a sustainable but otherwise largely static economy. Innovation will still be important. As we exhaust some resources, we will need replacements.

Consider our reliance on fossil fuels, rare earths needed in electronics, raw materials needed for steel and concrete, and fresh water. As these resources diminish, research and innovation will be needed to compensate for their scarcity.

How will our profession—computer science—contribute to such a situation? First, computers and their software are becoming more critical to successful research in most fields. “Computational-X” for many values of “X” is heard more often in halls of academe and industry. So is machine learning for many applications, although it is not the panacea that some enthusiasts might lead people to think. Our ability to accumulate scientific data and make it more usefully relevant is improving and that will contribute to the discovery of solutions to resource scarcity and substitution. We should not forget the potential for increasing optimality in manufacturing and the associated supply chains. Someday, quantum computing may supply near real-time optimal solutions to scheduling, resource allocation, and other similar and potentially large-scale problems.

A key question is whether a non-growing economy is feasible and habitable. For centuries the notion of “a growing business” had high value as an objective. A growing population contributed to that growth with its implicit increase in consumption. There is evidence, however, that this dynamic is changing. With time, we are seeing birth rates declining worldwide as non-agricultural economies demand less manual work, much of which is now done with machines. Robotics may also contribute to the reduction

in demand for manual labor. It is certainly possible to imagine a static population with increasing consumption and demand satisfied by automation. But machines consume resources too for their manufacture and operation so that, too, may reach some limits.

The questions in my mind remain: Is a static economy feasible? Desirable? Habitable? Much deeper analysis is needed. One also wonders what the transition to such an economy would look like. I wonder the same for the transition from fossil fuel to electricity for vehicle propulsion. At some point you won’t have enough fuel left in the tank to get to the nearest gas station for a refill. How will that process unfold and does it teach us anything about other scenarios in which what was once a vital resource is gone. Wood burning was replaced with alternative fuels—gas, coal, electricity—for cooking. All of those means were and still are concurrently in use although wood and coal are out in the long tail at this point with some exceptions in places where alternatives are too expensive to supply. At least wood is renewable, but excessive use can lead to disaster as seen in Haiti and in examples from Jared Diamond’s book *Collapse*.^a

It remains for us to explore these potential transitions out of concern for resource scarcity, hazards of global warming, and the evident slowing of human population growth. I look forward to an ensuing discussion! □

a J. Diamond. *Collapse*. Viking Press, 2004.

Vinton G. Cerf is vice president and Chief Internet Evangelist at Google. He served as ACM president from 2012–2014.

Copyright held by author/owner.

The *Communications* Web site, <http://cacm.acm.org>, features more than a dozen bloggers in the BLOG@CACM community. In each issue of *Communications*, we'll publish selected posts or excerpts.

twitter

Follow us on Twitter at <http://twitter.com/blogCACM>

DOI:10.1145/3449047

<http://cacm.acm.org/blogs/blog-cacm>

The SolarWinds Hack, and a Grand Challenge for CS Education

John Arquilla analyzes the latest in a long line of cyber intrusions, while Mark Guzdial considers how to prepare CS students for professional decision making.



**John Arquilla
From Solar Sunrise
to SolarWinds**

December 21, 2020

<http://bit.ly/2YD1e9c>

For all its breadth, depth, and skillful insertion via the supply chain, the latest hack of critical departments of the U.S. government—and of many leading corporations from around the world—should come as no surprise. Twenty-two years ago, as American forces were readying to strike Iraq for violations of an agreed-upon U.N. weapons inspection regime, deep intrusions into sensitive military information systems were detected. Enough material was accessed that, if printed out, it would have made a stack over 500 feet tall. The investigation into this hack, code-named “Solar Sunrise,” unearthed a group of teenagers, two in Northern California, one in Canada, and a young Israeli computer wizard, Ehud Tenenbaum.

The youth of the miscreants, and their lack of connection to a hostile

power, led to a somewhat dismissive attitude toward this sort of cyber threat. The absence of a sense of urgency about the problem was noted in a study of the matter undertaken by the National Academy of Sciences the following year, 1999, at a time when yet another very grave series of intrusions into American defense information systems—this time seemingly by Russians—was occurring. The effort to detect, track, and then deter further hacks was code-named “Moonlight Maze,” an

“The SolarWinds affair is simply another incident in a long pattern of intrusions.”

investigation that revealed the intrusions had been ongoing or at least three years before having been spotted.

It also took about three years to catch on to an apparent Chinese effort that had been cyber-snooping in sensitive American national security systems as well. That was back in 2003, when the “Titan Rain” forensic investigation got under way in earnest. Ever since, the Chinese efforts, led it is thought by their elite Unit 61398, have focused more on industrial and commercial intellectual property theft rather than on specifically military matters, to the tune of what is thought to be hundreds of billions of dollars worth of cutting-edge information.

The SolarWinds affair is simply another incident in a long pattern of intrusions. Yes, the angle of inserting malware in software specifically designed to enhance security is a creative touch. But we in the defense world have long been aware of this means of insertion. Indeed, I had graduate students at the military school where

I teach working on exactly this sort of problem many years ago. And the Stuxnet hack of the Iranian nuclear program a decade ago operated through the supply chain as well.

Why, then, this worst-ever hack? The National Academy study from 1999 put the matter well when it focused on an organizational culture, especially in the military, that tended to downplay thinking and planning for defense. To this I would add that, when conceiving of defense, too much reliance is placed on firewalls and anti-viral software designed to keep intruders out. These are Maginot Lines. Instead, the right approach is to “imagine no lines,” to think in terms of aggressors who will always find a way in. By cultivating a mind-set emphasizing this inevitability, those charged with protecting our cyberspace will find that innovative defensive practices will arise more readily.

For example, replacing the current faith in triple-belt firewalls with the ubiquitous use of very strong encryption will improve cyber defenses immeasurably. For it should be obvious by now that data at rest is data at risk. And beyond more and better use of encryption, sensitive data should also be kept moving. In the Cloud, even around in the Fog (populated by “edge devices” such as routers and switches that provide entry into enterprise or provider networks), the combination of strong crypto and cloud and edge computing will frustrate even the best cyber spies.

What is to be done now? Aside from fundamentally shifting the emphasis away from “static” cyber defenses such as fortified firewalls and anti-viral software that find it difficult to detect the latest advances in malware, it is crucially important to take full advantage of the opportunity the SolarWinds hack has provided to scour all information systems for any signs of delayed-action devices—designed not for spying, but rather for disrupting or distorting data flows in time of war. Military and business information systems should both get a clean bill of health; that is, test negative for signs of “cybotage,” before shifting to a new security regime based on strong codes and regular movement of data.

Such a scrubbing makes for a tall order. But unless action is undertaken now, the risk will grow that the next So-

larWinds-like event will come in a time of crisis or conflict, when lives are at stake and the price of complacency will be paid with the blood of soldiers frantically trying to access vital systems that no longer work.



Mark Guzdial
Teaching Critical Computing is a Grand Challenge for the Whole CS Curriculum

December 28, 2020

<https://bit.ly/3oASM4U>

The October 2020 issue of *Communications* had an education column by Amy Ko and her students, “It is time for more critical CS education” (see the paper at <https://bit.ly/3jfnhw3>). I had been looking forward to this paper since I saw Ko give the keynote talk on this topic at the Koli Calling conference in 2019, “21st Century Grand Challenges in Computing Education” (see the YouTube video at <https://bit.ly/39GODrN>). The authors argue that computing is so pervasive and critical to modern society that we need to prepare students to make decisions as professionals that are careful with the power that they are wielding. We must be teaching students that:

- ▶ Computing has limits.
- ▶ Data has limits.
- ▶ CS has responsibility.

I highlight here one particular paragraph in the paper:

Realizing a more critical CS education requires more than just teachers: it also requires CS education research. How do we teach the limits of computing in a way that transfers to workplaces? How can we convince students they are responsible for what they create? How can we make visible the immense power and potential for data harm, when at first glance it appears to be so inert? How can education create pathways to organizations that meaningfully prioritize social good in the face of rising salaries at companies that do not?

I strongly agree that we need CS education researchers to figure out how to achieve these goals, because we don’t know how right now. I also agree that we need more than “just teachers.” We need ALL CS teachers. You don’t meet a grand challenge with a handful of education researchers. A grand challenge requires a broad and pervasive response. We can use research from other-than-CS sources to identify the

issues in meeting the challenges in Ko et al.’s paper.

A significant risk of teaching students about critical computing is the risk of buoying confidence without imparting knowledge or changing behavior. There are questions about the effectiveness of ethics education, like this study in business (<https://bit.ly/36BtGN8>). Some studies of financial literacy education showed that students leave the course with greater confidence in their ability to make decisions, but without enough knowledge to actually make better decisions (see this study at <https://bit.ly/3jekweG>, and this study at <http://bit.ly/39GRqRX>). The concern is that we may give CS students the confidence that they know how to make critical decisions about computing, when they actually do not know enough to make those decisions or they don’t use the knowledge that they have effectively.

We cannot solve a grand challenge with a single course, either. Erin Cech is a sociologist who studies engineering education. She writes (see the paper at <https://bit.ly/3cy9SOv>) that we can’t get past the “culture of disengagement” unless we send a consistent message across the entire curriculum. A single “ethics” course sends the message that ethics is a one-shot deal, a box that you tick. Learning sciences research suggests that getting students to apply their knowledge in outside-the-classroom situations (the challenge of “transfer”) requires an approach that helps students connect the knowledge to several situations. If we want students to engage with ethical decision making, it has to be a message sent throughout the curriculum.

We need to prepare our students’ to have a critical perspective on computing. It is a research challenge, but it is also a challenge of will. We have to decide to meet this challenge as a field, not just with a course.

Thanks to Michael Kirkpatrick for pointing me to the Cech paper.

John Arquilla is Distinguished Professor of Defense Analysis at the U.S. Naval Postgraduate School. From 2005–2010, he served as Director of the Department of Defense Information Operations Research Center. The views expressed are his alone. **Mark Guzdial** is professor of electrical engineering and computer science in the College of Engineering, and professor of information in the School of Information, of the University of Michigan.

© 2021 ACM 0001-0782/21/4 \$15.00

A New Journal from ACM

Co-published with SAGE



Collective Intelligence, co-published by ACM and SAGE, with the collaboration of Nesta, is a global, peer-reviewed, open access journal devoted to advancing the theoretical and empirical understanding of collective performance in diverse systems, such as:

- human organizations
- hybrid AI-human teams
- computer networks
- adaptive matter
- cellular systems
- neural circuits
- animal societies
- nanobot swarms

The journal embraces a policy of creative rigor and encourages a broad-minded approach to collective performance. It welcomes perspectives that emphasize traditional views of intelligence as well as optimality, satisficing, robustness, adaptability, and wisdom.

Accepted articles will be available for free online under a Creative Commons license. Thanks to a generous sponsorship from Nesta, Article Processing Charges will be waived in the first year of publication.

For more information and to submit your work,
please visit <https://colint.acm.org>



Association for
Computing Machinery



The Best of NLP

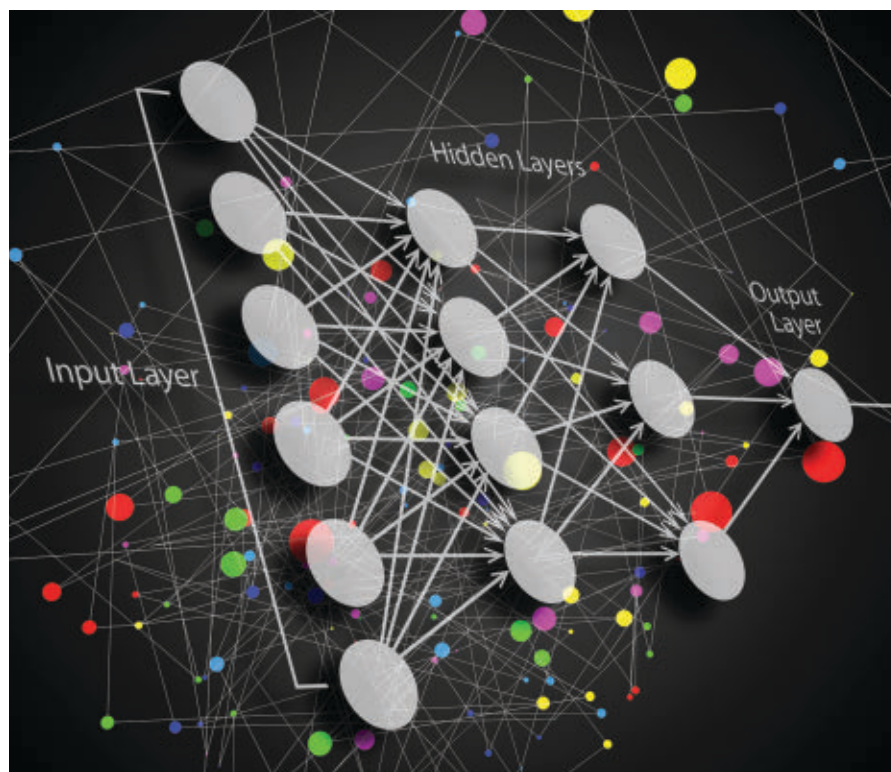
Natural language processing delves more deeply into its knowledge gap.

WHEN IT WAS released by Google just a few years ago, a deep-learning model called BERT demonstrated a major step forward in natural language processing (NLP). BERT's core structure, based on a type of neural network known as a Transformer, has become the underpinning for a range of NLP applications, from completing search queries and user-written sentences to language translation.

The models even score well on benchmarks intended to test understanding at a high school level, such as Large-scale ReAding Comprehension (RACE) developed at Carnegie Mellon University. In doing so, they have become marketing tools in the artificial intelligence (AI) gold rush. At Nvidia's annual technology conference, president and CEO Jen-Hsun Huang used RACE to claim high performance for his company's implementation of BERT.

"The average human scored 73%. Expert humans score 95%. Nvidia's Megatron-BERT scored 91%," Huang said, adding, "Facebook AI Research developed a Transformer-based chatbot with knowledge, personality, and empathy that half of the users tested actually preferred [over humans]."

Performance stepped up another



notch with the release of GPT-3 in summer 2020, the latest iteration of a series of language models developed by the company OpenAI. Sporting 175 billion trainable parameters, GPT-3 is 500 times larger than BERT's biggest version.

Size has given GPT-3 seemingly impressive abilities. Whereas most other

Transformer-based systems need a training sequence that "fine-tunes" the last few layers of the deep neural-network (DNN) pipeline to fit a specific application, such as language translation, OpenAI promises GPT-3 can dispense with the need for extensive fine-tuning because of the sheer size of its core training set.

Tests have demonstrated the ability of GPT-3 to construct lengthy essays in response to brief prompts. Yet the huge system has flaws that are easy to show. Questions to GPT-3 often can yield answers of almost nightmarish surrealism, claiming in one case that blades of grass have eyes, or in other situations that a horse has four eyes. OpenAI's own research team questioned the limits of huge models trained purely for language modeling in a paper published shortly after the release of GPT-3.

The key to the performance of these language models seems to come down to their ability to capture and organize sometimes contradictory information mined from enormous collections of text that include sources such as Wikipedia and the social-media site Reddit. Early approaches used word embedding, in which each discrete word is converted to a numeric vector using a clustering algorithm. Words that most commonly surround it in the corpus used for training determine the vector's values. But these approaches hit problems because they could not disambiguate words with multiple meanings.

The networks inside BERT take the flexible meanings of words into account. They use multiple layers of neural-network constructions called Transformers to assign vectors not to separate words, but to words and subwords in different contexts that the model finds as it scans the training set.

Though Transformers associate words and their stems with different contexts, what remains far from clear is what relationships between words and context they actually learn. This uncertainty has spawned what University of Massachusetts Lowell assistant professor Anna Rumshisky and colleagues termed "BERTology." BERT is a particular focus in research like this because its source code is available, whereas the much larger GPT-3 is only accessible through an API.

Closer inspection of their responses shows what these systems clearly lack is any understanding of how the world works, which is vital for many of the more advanced applications into which they are beginning to be pushed. In practice, they mostly make associations based on the proximity of words

in the training material; as a result, Transformer-based models often get basic information wrong.

For example, Ph.D. student Bill Yuchen Lin and coworkers in Xiang Ren's group at the University of Southern California (USC) developed a set of tests to probe language models' ability to give sensible answers to questions about numbers. BERT claims a bird has twice the probability of having four legs rather than two. It also can give contradictory answers. Though BERT will put a high confidence on a car having four wheels, if the statement is qualified to "round wheels," the model claims it is more likely to sport just two.

Toxicity and unwanted biases are further issues for language models, particularly when they are integrated into chatbots that might be used for emotional support: they readily regurgitate offensive statements and make associations that tend to reinforce prejudices. Work by Yejin Choi and colleagues at the Allen Institute for AI has indicated a major problem lies in subtle cues in the large text bases used for training that can include sources like Reddit. However, even training just on the more-heavily-policed Wikipedia show issues.

"Sanitizing the content will be highly desirable, but it might not be entirely possible due to the subtleties of potentially toxic language," Choi says.

One way to improve the quality of results is to give language models a better understanding of how the world works

Choi points to the issue that training on conventional text suffers from reporting bias; even encyclopedic sources do not describe much of how the world around us works.

by training them on "commonsense" concepts. This cannot be achieved by simply giving them bigger training sets. Choi points to the issue that training on conventional text suffers from reporting bias: even encyclopedic sources do not describe much of how the world around us works. Even worse, sources such as news, which supports much of the content of Reddit and Wikipedia, express exceptions more often than the norm. Much of the background knowledge is simply assumed by humans; to teach machines, this background calls for other sources.

One possible source of commonsense knowledge is a knowledge base, which needs to be built by hand. One existing source that some teams have used is ConceptNet, but it is far from comprehensive.

"We need knowledge of why and how," Choi notes, whereas the majority of elements in ConceptNet typically describe "is a" or "is a part of" relationships. To obtain the information needed, the group crowdsourced the information they wanted for their own Atomic knowledge base. They opted to build a new knowledge base rather than extend ConceptNet, partly because it focused the fine-tuning on aspects of behavior and motivation without potentially extraneous information, but also because Atomic is expressed in natural-language form, so the knowledge can more easily be processed by BERT. ConceptNet's symbolic representations need to be converted to natural language form using templates.

However, it remains unclear whether the Transformer neural-network design itself provides an appropriate structure for representing the knowledge it attempts to store. Says Antoine Bosselut, postdoctoral researcher at Stanford University, "It's one of the most interesting questions to answer in this space. We don't yet know exactly how the commonsense knowledge gets encoded. And we don't know how linguistic properties get encoded."

To improve the abilities of language models, Tetsuya Nasukawa, a senior member of the technical staff at IBM Research in Japan, says he and his colleagues took inspiration from the way images and language are used together to teach children, when creating

their visual concept naming (VCN) system. This uses images and text from social media to link objects to the words often used to describe them, on the basis that different cultures and nations may use quite different terms to refer to the same thing, and which are not captured in conventional training based on text alone. “We believe it’s essential to handle non-textual information such as positions, shapes, and colors by using visual information,” he says.

Another approach, which has been used by Ren’s group, is to take an existing handbuilt knowledge base and couple it to a Transformer, rather than trying to teach the language model common sense. KagNet fine-tunes a BERT implementation in conjunction with a second neural network that encodes information stored in the ConceptNet knowledge base.

An issue with linking Transformers to other forms of AI model is that it is not yet clear how to make them cooperate in the most efficient manner. In the USC work, the KagNet does not add much in terms of accuracy compared to a fine-tuned language model working on its own. As well the relative sparsity of information in the knowledge base, Lin says the knowledge-fusing method may not go deep enough to make good connections. A further issue common to much work on language models is that it is not easy to determine why a language model provides the answer it does. “Does the model really answer the question for the right reasons? The current evaluation protocol may not be enough to show the power of symbolic reasoning,” Lin says.

Nasukawa says work in visual question answering, in which a system has to answer a textual question about the content of an image, has met with similar issues. He says the most productive route that has emerged so far is to tune the second architecture for a specific application, rather than trying to fine-tune something more generic in the way language models currently work. A more sophisticated general-purpose structure that can be used across many applications has not yet emerged for applications that need understanding of how the world works. In the meantime, Transformers may yield more

Another approach is to take an existing handbuilt knowledge base and couple it to a Transformer, rather than trying to teach the language model common sense.

surprises as they continue to scale up.

“Each time, the added scale gives us new capabilities to let us test new assumptions,” Bosselut says. “As much as many people think we are going too far down this path, the truth is that the next iteration of language modeling could open a new set of capabilities that the current generation doesn’t have. This is a great thing about NLP: there does seem to be an openness to diverse perspectives.” **C**

Further Reading

Rogers, A., Kovaleva, O., and Rumshisky, A. *A Primer in BERTology: What we know about how BERT works* arXiv:2002.12327 (2020) <https://arxiv.org/abs/2002.12327>

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL) (2019).*

Lin, B.Y., Chen, X., Chen, J., and Ren, X. *KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019)*

Muraoka, M., Nasukawa, T., Raymond, R., and Bhattacharjee, B. *Visual Concept Naming: Discovering Well-Recognized Textual Expressions of Visual Concepts, Proceedings of The Web Conference (2020)*

Chris Edwards is a Surrey, U.K.-based writer who reports on electronics, IT, and synthetic biology.

© 2021 ACM 0001-0782/21/4 \$15.00

ACM Member News

ON THE TRAIL OF HUMAN-AWARE AI



Subbarao Kambhampati is a professor in the Department of Computer Science and Engineering at

Arizona State University in Tempe, AZ.

He received his undergraduate degree in Electrical Engineering and Electronics from the Indian Institute of Technology in Madras, India. “I worked on speech recognition systems for my bachelor’s thesis, and that got me interested in AI (artificial intelligence) in general,” Kambhampati says. To further immerse himself in AI, he began to study computer science.

Kambhampati went on to earn his master’s degree and Ph.D.—both in computer science—from the University of Maryland, in College Park, MD.

After a brief post-doctoral stint at Stanford University, Kambhampati joined the faculty of Arizona State University in 1991, where he has remained ever since.

Kambhampati’s main research interests include automated planning and decision making. His current research focus is on human-aware AI systems and how such systems plan their behavior, particularly when there are humans in the loop.

“My main interest is in getting AI systems to interact with humans in fluid ways,” Kambhampati says. His goal is to have AI systems operate unobtrusively in the background, in a benign way.

“The headline for AI should not be sinister,” Kambhampati says, “but that it helps some old woman cross the road.”

Kambhampati has served as president of the Association for the Advancement of Artificial Intelligence, and he works to get people to understand the real issues around AI, to dispel the hype. One means to that end is a column on the impact of AI that he writes for *The Hill*, a news website based in Washington, D.C.

—John Delaney

Deep Learning Speeds MRI Scans

Machine intelligence significantly reduces the time needed for an MRI scan, which can help reduce patient anxiety.

SINCE ITS INVENTION in the 1970s, magnetic resonance imaging (MRI) has opened up a window onto the world beneath our skin. By exploiting the way the nuclei of hydrogen atoms in water and fat molecules resonate in a strong magnetic field, MRI can generate high-contrast three-dimensional images of soft body tissues, joints, and bones. MRI allows clinicians to see evidence of injury and disease within the body, ranging from torn muscle to damaged cartilage, ligaments, and tendons, as well as tumors or other disease lesions within major organs, and blood-flow blockages in the brain, all without the ionizing radiation of the X-rays used in computed tomography (CT) scans.

There is, however, a considerable usability problem with the MRI scanner as we currently know it: the technology takes far too long to acquire images, forcing patients to lie still in the confined maw of a massive magnet for up to an hour. With the observable world reduced to a halo of grayish plastic just inches from one's nose, it is a particularly tough experience for those suffering from claustrophobia. It can be disturbingly noisy, too: the scanner's magnetic components can rattle at 110 decibels or more when energized.

"It can take typically three or four minutes to acquire each magnetic resonance image, and if you're lying on the table for 30 or 40 minutes, or sometimes even an hour, depending on the type of exam, it gets hard to lie still for that entire time. It's uncomfortable for the patient, and especially so if they're a child," says Michael Recht, Louis Marx Professor and chairman of the department of radiology at NYU Langone Health.

Help is now at hand, and from an unlikely quarter. Facebook, the Palo



Alto, CA-based online social network, has teamed up with radiologists at New York University's Langone Medical Center in Manhattan to develop an artificial intelligence (AI)-based imaging accelerator for MRI scanners.

What Facebook and NYU Langone have developed is a deep learning neural network (DNN) that allows MRI scans to be performed many times faster. Called FastMRI, the DNN, has been trained to generate MRI images using far less magnetic resonance data than before—and that sparse data requirement significantly accelerates the scan time.

Pedal To The Metal

Acceleration is, in fact, a long-held quest in MRI, says Recht's colleague, Daniel Sodickson, a professor of radiology, neuroscience, and physiology at NYU Langone Health and a specialist in biomedical imaging. He should

know: in 1996, he developed an early MRI speed-up technique called parallel imaging. To understand how this works, consider the way an MRI scanner operates in general:

- ▶ When the patient is placed in the magnetic field of the scanner's main magnet, hydrogen nuclei (protons) in the body's water and fat molecules line up with that magnetic field;

- ▶ Three orthogonal electromagnetic coils, in the x, y and z planes, project pulsed, spatially varying magnetic field gradients into the body, making the protons momentarily change their polarization direction;

- ▶ After each pulse projection subsides, the protons relax and line up once again with the main magnet, revealing their location by emitting distinctive radio frequencies that are picked up by a (receiver coil) detector;

- ▶ The frequency and phase of the detected signal allows the MRI soft-

ware to map the location of water and fat in the body, and to mathematically derive high contrast images from that data.

Sodickson realized the more detectors an MRI scanner could use at once, the faster an image could be computed from frequency and phase data—known as *k-space* data, as it is *not* image data in the conventional sense of being an array of pixels.

“Parallel imaging—gathering data in lots of different detectors arranged around the body—at least doubled our speed,” he says.

To speed the process further, the industry has tried a technique called compressed sensing, in which algorithms inform the array of detectors which *k-space* data they can *most probably* ignore. “It’s almost like pre-compressing an image with JPEG,” says Sodickson.

However, it is far from perfect: compressed sensing can lead to blocky, blurred image artifacts that might confound diagnosis. What is needed is a way to learn, with much higher accuracy, which *k-space* data does not need to be collected by the detectors. An MRI scanner, says Recht, might project 256 magnetic gradient signals into the body to give different *k-space* “views” of the area under scrutiny. However, because many of the signals overlap and some view angles might be unnecessary, it is highly likely many projections might be redundant and do not need to be taken.

“It’s just hard to say in advance which projections you can skip and so accelerate the scan. With deep learning, we can learn which ones you can skip and still produce the entire image,” says Sodickson.

It was in 2018, while the NYU Langone team were puzzling over this issue of *k-space* data “undersampling,” as it is called, and making their own preliminary experiments with acceleration by deep learning, that Sodickson discovered the Facebook AI Research (FAIR) lab was actively seeking projects in the “AI-for-good” arena, on which they hoped to have a positive societal impact.

A Battle Worth Fighting

When Sodickson and his colleagues told Facebook precisely what was need-

“FastMRI uses artificial intelligence to create images from the *k-space* data, and we’ve been able to train the AI to create accurate images from undersampled *k-space* data.”

ed, the Californians’ ears pricked up, he recalls: “They said, ‘wait a second, you want to reconstruct an image from not enough data? But you don’t want that image to just be plausible, you want it to actually be true to that patient? Now *that’s* an interesting AI problem.”

As a result, Facebook AI Research and NYU Langone agreed to jointly develop a deep-learning-based faster MRI system—and also that the project would be open sourced on Microsoft’s GitHub platform. Unlike neural networks trained to recognize an image or a voice from input data, the researchers had to craft a deep learning (DL) network capable of *generating* an accurate, diagnostic-quality image from an undersampled subset of an MRI scanner’s acquired *k-space* frequency and phase data.

Initially, Facebook took NYU Langone’s anonymized and open-sourced knee MRI dataset—which comprises the *k-space* projection data from 1,200 scans of 108 patients’ knees, and the full images the MRI software resolved—and coded up a standard-issue DL model that could learn the relationship between them. “But they got lousy results,” says Sodickson.

What they needed, he says, was a far more nuanced network informed by the physics of magnetic resonance. They then built a special type of DL model, called a variational network, that did not simply undergo blind ML training with *k-space* and image data alone: it also was trained with key information about the physics of the

scanner, including mapping variations in the way receiver coil sensitivities changed across detector arrays.

To test the idea, the joint team trained its network for 155 hours using eight cloud-based GPUs, and found their new, scanner-physics-aware approach made all the difference. They found the network was able to shed three-quarters of the raw *k-space* data, and still allow their AI model to generate diagnostic-quality images with an almost fourfold acceleration, the Facebook/Langone team reported in the December 2020 edition of the *American Journal of Roentgenology*.

Better still, the images of the accelerated knee scans were judged by a jury of six senior radiologists to be of better quality than the standard-speed images. Also, in early as-yet-unreported undersampled tests on MRIs of the brain, it looks like variational DL-based scans can be accelerated between six to eight times, says Recht, while Sodickson is predicting a 10-fold improvement for some types of abdominal scan. “So if that took an hour before, it would now just be six minutes in the scanner, or if it was ten minutes, it’d now just be one minute,” Sodickson says.

In getting faster, better images from less data, their result might seem counterintuitive, if not downright magical. Yet Anuroop Sriram, a senior Facebook AI research engineer on the FastMRI project, cautions it is important to remember the scanner is not simply sampling pixels like some kind of camera, but is capturing something quite abstract: raw frequency and phase data.

The traditional way of turning *k-space* data into a readable scan is to apply a mathematical process called an inverse Fourier transform, which translates it from the frequency domain to a spatially resolved image. “But if you try to use that process on less than a full scan of *k-space* data, you don’t end up with a useful image,” says Sriram.

“Our FastMRI approach is creating images in a completely new way: rather than using that mathematical process, FastMRI uses artificial intelligence to create images from the *k-space* data, and we’ve been able to train the AI to create accurate images from undersampled *k-space* data.”

Nafissa Yakubova, Facebook’s AI

program manager, believes the lab has hit its target of making a societal impact. “We’ve advanced AI to address this problem, and done so in a way that could actually one day be used in medical practice, benefiting patients, clinics, and communities,” she says.

To do that, says Recht, the Langone team is beginning a multihospital study in collaboration with market-leading MRI scanner vendors Siemens, General Electric, and Philips Healthcare. The overarching aim of the study is not only proving the Fast-MRI DL is generalizable across musculoskeletal, knee, brain and abdomen scans, but also across multiple vendors’ scanners. “Our goal is to get this as fast as possible, to as many companies as possible, so that they can make this available to patients everywhere,” says Sodickson.

Others are on the FastMRI group’s trail, with machine learning researchers at Imperial College London, the Korea Advanced Institute of Science & Technology (KAIST), Stanford University, and China’s Shenzhen Institute of Advanced Technology all independently researching their own deep learning-based methodologies for MRI acceleration.

“Deep learning is much better than the traditional parallel imaging and

“Deep Learning is much better than the traditional parallel imaging and compressed sensing approaches.”

compressed sensing approaches. Those classical approaches are usually based on top-down models, so if the model fails in a real acquisition scenario, image degradation is unavoidable,” says Jong Chul Ye, a signal processing and ML researcher at KAIST in Daejeon, South Korea.

The challenge now, Ye says, is to move MRI acceleration from the supervised learning the Facebook/NYU team used to train its variational model to more efficient unsupervised models. “Many groups in the imaging community, including mine, are now working on unsupervised learning approaches. This area is still quite an open one, and it’s one that’s going to need a lot of machine learning know-how.”

Further Reading

A large-scale dataset of both raw MRI measurements and clinical MRI images <https://github.com/facebookresearch/fastMRI/>

Cha, E., Oh, G., and Ye, J. C. Geometric Approaches to Increase the Expressivity of Deep Neural Networks for MR Reconstruction *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1292-1305, Oct. 2020
doi: 10.1109/JSTSP.2020.2982777

Recht, M. P., Sodickson, D. K., Knoll, F., Yakubova, N., Zitnick, C. L., et al Using Deep Learning to Accelerate Knee MRI at 3T: Results of an Interchangeability Study, *American Journal of Roentgenology*, December 2020, Vol. 215, No.6, pp.1421-1429
doi.org/10.2214/AJR.20.23313

Sodickson, D. K., and Manning, W.J. Simultaneous Acquisition of Spatial Harmonics (SMASH): Fast Imaging with Radiofrequency Coil Arrays, *Magnetic Resonance in Medicine*, October 1997, Vol 38 (4), pp.591-603
<https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.1910380414>

Breaking the MRI Sound Barrier, General Electric
<https://www.gehealthcare.co.uk/feature-article/breaking-the-mri-sound-barrier>

Paul Marks is a technology journalist, writer, and editor based in London, U.K.

© 2021 ACM 0001-0782/21/4 \$15.00

CACM News

Eyes on the Skies

The prospect of drones delivering goods to us just 30 minutes after we click “buy now” online is certainly a compelling one, especially so in the era of Covid-19 lockdowns.

Yet delivery drones present “serious privacy concerns” that need addressing, warn security researchers at the Indian Institute of Science (IISc).

At issue, says lead researcher Vinod Ganapathy, director of computer systems security at IISc, is that delivery drones are much more than cargo-carrying flying machines: they are airborne, wirelessly connected, location-aware computer platforms peppered with potentially invasive arrays of sensors.

This means as a parcel-laden drone flies to its destination,

its sensors might also be able to capture a great deal of data on households and their inhabitants, the IISc team says. A drone’s cameras could capture images or video of people in a house, as well as vehicles and their license plates outside, and laser-ranging LiDAR sensors could acquire data about buildings and outbuildings. The IISc team’s concern is that data acquired in that manner could be used by a logistics firm or marketed to data brokers to target households with advertising.

Alongside colleagues Rakesh Rajan Beck and Abhishek Vijeev, Ganapathy has developed a potential solution: a software framework that drone operators can adopt to conform with local

privacy laws. Called Privaros, the privacy-enforcing software framework is designed to work with the middleware at the heart of most drones: the real-time version of Willow Garage’s Robot Operating System (ROS2).

Privaros works well, the team reports, because its privacy rules harness ROS2 procedures similar to those that allow drones to obey national flight rules. Privaros is globally portable, says Ganapathy, and can be adapted to regulations that may be developed by the U.S. Federal Aviation Administration or the European Aviation Safety Agency.

Nirupam Roy, a delivery drone security researcher at the University of Maryland at College Park, is impressed. “The Privaros team have identified

this privacy and security concern, and proposed a practical framework for privacy-compliant navigation of delivery drones. Delivery drones are now a reality, so it is definitely very timely research, and it is a thorough implementation they have built,” says Roy.

Ganapathy is undaunted, and hopes other developers will help them make drone deliveries a success, privacy-wise. Says Ganapathy, “These are early days in the drone privacy space, and Privaros is an early technology that’s designed to help. We call on the community to build upon Privaros to address this important problem.”

—Paul Marks is a technology journalist, writer, and editor based in London, U.K.

The Worsening State of Ransomware

Sophisticated, dangerous ransomware is the new normal ... and there is no simple fix.

FEW THINGS ELICIT terror quite like switching on a computer and viewing a message that all its files and data are locked up and unavailable to access. Yet, as society wades deeper into digital technology, this is an increasingly common scenario. Ransomware, which encrypts data so cybercriminals can extract a payment for its safe return, has become increasingly common—and costly. A 2019 report from security vendor Emisoft pegged the annual cost of ransomware in excess of \$7.5 billion in the U.S. alone.¹

“Individuals, businesses, hospitals, universities and government have all fallen victim to attacks,” says Chris Hinkley, head of the Threat Resistance Unit (TRU) research team for security firm Armor. In a worst-case scenario, ransoms can run into the tens of millions of dollars and close down an organization’s operations entirely. It has forced hospitals to redirect patients to other facilities, disrupted emergency services, and shut down businesses.

The problem is growing worse, despite the development of new and more advanced ways to battle it, including the use of behavioral analytics and artificial intelligence (AI). “Cybergangs use different cryptographic algorithms and they distribute software that is remarkably sophisticated and difficult to detect,” Hinkley says. “Today, there is almost no barrier to entry and the damage that’s inflicted is enormous.”

Money for Nothing

The origins of modern ransomware can be traced to September 2013. Then, a fairly rudimentary form of malware, CryptoLocker, introduced a new and disturbing threat: when a person clicked a malicious email link or



opened an infected file, a Trojan Horse began encrypting all the files on a computer. Once the process was complete, crooks demanded a cryptocurrency payment, usually a few hundred dollars, to unlock the data. If the person didn’t pay in cybercurrency, the perpetrator deleted the private key needed to decrypt the data and it was lost permanently.

Today, a dizzying array of ransomware exists, with each variation developed by different cybergangs. Once they reside on a computer, the likes of Dharma, Maze, Ryuk, Petya, Sodinokibi, Lazarus, and Lockbit unleash malware that spreads across systems and networks—until the crooks decide to pull the trigger. Making matters worse, some cybergangs sell ransomware kits for as little as a few hundred dollars (or via a subscription that may run as low as \$50 to \$100 per month). These “customers,” who have zero coding skills or software expertise, take advantage of a ransomware-

as-a-service (RaaS) model to gain sophisticated capabilities, says Keith Mularski, a former FBI agent and now managing director of the cybersecurity practice at Ernst & Young.

According to security firm Sophos, 51% of organizations it sampled globally found themselves the targets of ransomware attacks in 2019. The crooks succeeded in encrypting data in 73% of these attacks. Just over a quarter of these organizations paid the ransom, or their insurance companies forked over the cash. For instance, University Hospital of New Jersey paid a \$670,000 ransom in October 2020 after a group called SunCrypt captured 240GB of its data. A more catastrophic outcome occurred in July 2019, when Portland, OR-based PM Consultants, a managed services provider (MSP) for dental practices, was hit with ransomware; customers could not access key files or data for months, and the firm shut down.⁵

Not surprisingly, dozens of major ransomware gangs now exist worldwide, including in Russia, Eastern Europe, and North Korea. Incredibly, many of these operations look and function like authentic businesses. “They rent office space, they have development teams, data architecture teams, help desks, phone support, and people that negotiate ransoms with targets,” says Alexander Chaveriat, chief innovation officer at Tuik Security Group. “They buy server space all over the world using cryptocurrency, change servers as needed, and use virtual private networks and other tools to hide their location.”

Code Red

Although ransomware attacks vary, an episode begins when a computer executes an infected file. The malware usually downloads additional components that establish a connection to a Command and Control (C&C) server. This allows data to flow across the machines—including an IP address, geo-location data, and information about access permissions. This connection is referred to as a “call home” or “C2,” and it typically taps Port 80 and HTTP or Port 443 and HTTPS protocols. At some point, the crooks load an encryption key needed to lock the files onto the target computer.⁶

The encryption process ensues over days, weeks, or months, normally progressing through hard drives, attached drives, and network devices. The C&C server decrypts files as they are needed. Along the way, crooks place a ransom note in every folder that has encrypted files; they might also plant other types of malware on systems. During the final stage of an attack, the ransomware uninstalls itself, the thieves remove the encryption key from the infected system and the victim sees a ransom note on the computer screen.⁷

The mechanics of ransomware have advanced considerably over the last few years. Early assaults were largely automated and focused on infecting large numbers of computers. Demands of \$400 to \$1,000 were common, says John Shier, senior security advisor at Sophos. As patching and endpoint security have improved, ransomware has evolved. In many cases, cybergangs—sometimes with the sup-

port of nation-states—target specific businesses, hospitals, or cities. In fact, they frequently seek out organizations with cyber-insurance, which increases the odds they can cash in.

Consider Emotet, a ransomware “dropper” that lands on a system after a person clicks on a malicious e-mail link, executes an infected file, or clicks on a hijacked online ad that contains malicious code. This installs the initial Emotet malware on a computer. That malware, in turn, downloads scripts, macros, and code that pull data from address books, use password stuffing to break into other accounts, and install spyware. Emotet components hide in sandboxes, slip into cloud containers, and escape detection by firewalls as a result of encrypted communications channels.

Along the way, different cybergangs and various forms of malware go to work. This includes banking trojans like Dridex and Trickbot, “middle-stage infectors that steal credentials so that criminals can perpetrate some type of financial crime,” Shier says. “Once a group is finished stealing credentials, they hand things over to a ransomware operator, who encrypts the machine and demands the payment.” Sophos identified upward of 700 unique Emotet binaries appearing per day in 2019, something that makes conventional signature-based identification next to impossible. “What started as a monolithic code base, includ-

“Suddenly, you have people with limited ability using powerful software to discover, exfiltrate, and encrypt files. They wind up with many of the same capabilities that sophisticated cybercriminals have.”

ing a credential stealer, has become a highly modular payload that allows operators to mix and swap out components,” according to Shier.

Another common ransomware package, Dharma (previously known as CrySis), attacks small and medium-sized businesses. While the average ransomware demand is now \$191,000, according to Sophos, Dharma lands at a relatively low figure of \$8,620. “The ransomware crew that produced Dharma has put it in the hands of lower-skilled criminals,” Shier says. “Suddenly, you have people with limited ability using powerful software to discover, exfiltrate, and encrypt files. They wind up with many of the same powerful capabilities that sophisticated cybercriminals have at their disposal.”

Sophos has found that 85% of Dharma infections were associated with vulnerabilities in Remote Desktop Protocol (RDP), a proprietary Microsoft communications protocol that facilitates connections between corporate networks and remote computers. Vulnerable systems typically lack multi-factor authentication, so after paying a fee or buying a subscription, an affiliate obtains a menu-driven PowerShell script that establishes a connection to a business through RDP. The package includes a credential-stealing tool called Mimikatz, along with various other system utility tools.⁹

Methods to the Madness

Ransomware techniques continue to evolve. A good example is a program called Snatch, introduced in 2019. During the initial infection phase, the malware sets registry keys that are needed to run a particular file in Safe Mode. After planting the encryption program, it points the registry keys at it and then reboots the machine. Once the computer is in safe mode, with normal security tools switched off, it can encrypt files unimpeded. Other evasive techniques it uses include initiating attacks within virtual machines, and encrypting files in memory to avoid behavioral detection methods.¹⁰

Gangs also have begun encrypting backup systems, including cloud storage services such as Office 365 and Dropbox. Although 56% of the firms surveyed by Sophos regained control of their data through backups, that window appears

to be closing. “[Cybergangs] have realized that the ransom demand becomes powerless if you have a full backup set in place and you can revert to it,” Shier says. The gangs also are discovering ways to ratchet up the pressure. Beginning in November 2019, a group associated with Maze ransomware began copying data from targeted systems before encrypting it—something other groups have since copied. This can include human resources records, legal information, and intellectual property. Frequently, they post samples online to verify they hold these documents and data.

In May 2020, for instance, celebrity law firm Grubman Shire Neusekas & Sacs found itself in the crosshairs of an initial ransomware demand of \$21 million, Armor says. The ransomware gang responsible for the attack claimed it held thousands of documents, containing the private information of Lady Gaga, Nicki Minaj, Bruce Springsteen, LeBron James, Christina Aguilera, Mariah Carey, and others. When the law firm failed to respond to the ransom demands, the gang doubled the ransom to \$42 million. On July 10, the gang began auctioning the private data on the Dark Web for as much as \$1.5 million per cache.¹¹

The ability to capture financial data has other consequences. “Ransomware operators can use it to determine how much money an organization can afford to pay for a ransom,” says Chavriat. Not surprisingly, this can drive up the price of the ransom, while defusing any argument the business does not have the cash the bandits are demanding. “There have been cases where the thieves asked for the exact amount of money covered by the insurance and corporate policy. This indicates they have access to extracted data,” he says.

Ransomware attacks also are spreading to industrial control systems. In 2019, Norwegian aluminum manufacturer Norsk Hydro suffered an attack that forced the company to switch some operations to manual mode. The company reported total estimated losses from the incident exceeded US\$40 million. Now there is concern that ransomware will spread to connected Internet of Things (IoT) devices such as automobiles, home automation systems, and medical devices, Hinkley says.

Living in a world teeming with ransomware is a growing concern. It is impossible to know the full extent of the damage, because many victims don't report attacks.

Exiting the Maze

Living in a world teeming with ransomware is a growing concern. It is impossible to know the full extent of the damage because many victims don't report attacks. According to Sophos, 94% of organizations whose data was encrypted regained control of it by paying a ransom, or through backups. “It's in the best interest of gangs to ensure that people do get their data back. You're more likely to pay if you trust criminals to honor their end of the deal,” Shier says. Yet that came at an average cost of nearly \$1.5 million per incident, when downtime, people time, device cost, network cost, lost opportunity and the ransom paid are factored into the equation.¹³

For a number of reasons, including a lack of international extradition treaties, few ransomware gangs are ever brought to justice. Some, including the U.S. Treasury, have promoted the idea of making it illegal to pay a ransom, though the idea has not gained widespread support. One way the computing industry is fighting back is by taking down C&C servers. Last October, Microsoft disrupted an enormous hacking operation after it obtained a U.S. federal court order to disable the IP addresses associated with Trickbot's servers and worked closely with telecom providers to eradicate hackers.¹⁵

Cybersecurity experts do not see an end to ransomware anytime soon. Artificial intelligence, blockchain, and other technologies may improve

detection and protection—and employee training may improve detection—but every time there is an advance in defense, cybergangs find a new way to breach systems and extract ransoms. Says Mularski: “As cybersecurity has improved, ransomware gangs have continued to find the weakest links and exploit them. When people asked Willie Sutton why he robbed banks, he replied: ‘That's where the money is.’ Today, ransomware is where the money is.” ■

Endnotes

1. The State of Ransomware in the US: Report and Statistics 2019, *Emisoft*, <https://bit.ly/36qYkbf>
2. The State of Ransomware 2020, *Sophos*, <https://bit.ly/3dtIbb>
3. *Ibid.*
4. New Jersey hospital paid ransomware gang \$670K to prevent data leak, *BleepingComputer*, <https://bit.ly/3jXOWBj>
5. The new target that enables ransomware hackers to paralyze dozens of towns and businesses at once, *GCN*, <https://gcn.com/articles/2019/09/12/ransomware-msp.aspx>
6. “Ransomware: How an attack works,” *Sophos*, https://support.sophos.com/support/s/article/KB-000036277?language=en_US
7. *Ibid.*
8. Color by numbers: Inside a Dharma ransomware-as-a-service attack, *Sophos News*, <https://bit.ly/379rkGd>
9. *Ibid.*
10. The realities of ransomware: The evasion arms race, *Sophos News*, <https://bit.ly/2I188v0>
11. The Dark Market Report: The New Economy, *Armor*, <https://www.armor.com/resources/the-dark-market-report/>
12. The State of Ransomware 2020, *Sophos*, <https://bit.ly/3dtIbb>
13. *Ibid.*
14. Ransomware attacks are increasing at an unprecedented rate — and the US is now begging people not to pay ransoms, *Business Insider*, <https://bit.ly/3k590qh>
15. Microsoft takes down massive hacking operation that could have affected the election, *CNN Business*, <https://cnn.it/3dtLX0u>

Further Reading

Richardson, R. and North, M.M.

Ransomware: Evolution, Mitigation and Prevention, 2017, Faculty Publications, 4276, <https://digitalcommons.kennesaw.edu/facpubs/4276>

Kok, S.H., Abdullah, A., and Jhanjhi, N.Z.
Early detection of crypto-ransomware using pre-encryption detection algorithm, July 4, 2020, <https://doi.org/10.1016/j.jksuci.2020.06.012>

Hull, G., Henna, J., and Arief, B.
Ransomware deployment methods and analysis: views from a predictive model and human responses, *Crime Science*, February 2019, <https://link.springer.com/article/10.1186/s40163-019-0097-9>

Samuel Greengard is an author and journalist based in West Linn, OR, USA.

ACM ON A MISSION TO SOLVE TOMORROW.



Dear Colleague,

Without computing professionals like you, the world might not know the modern operating system, digital cryptography, or smartphone technology to name an obvious few.

For over 70 years, ACM has helped computing professionals be their most creative, connect to peers, and see what's next, and inspired them to advance the profession and make a positive impact.

We believe in constantly redefining what computing can and should do.

ACM offers the resources, access and tools to invent the future. No one has a larger global network of professional peers. No one has more exclusive content. No one presents more forward-looking events. Or confers more prestigious awards. Or provides a more comprehensive learning center.

Here are just some of the ways ACM Membership will support your professional growth and keep you informed of emerging trends and technologies:

- Subscription to ACM's flagship publication *Communications of the ACM*
- Online books, courses, and videos through the **ACM Learning Center**
- Discounts on registration fees to ACM Special Interest Group conferences
- Subscription savings on specialty magazines and research journals
- The opportunity to subscribe to the **ACM Digital Library**, the world's largest and most respected computing resource

Joining ACM means you dare to be the best computing professional you can be. It means you believe in advancing the computing profession as a force for good. And it means joining your peers in your commitment to solving tomorrow's challenges.

Sincerely,

A handwritten signature in black ink, appearing to read 'G. Kotsis'.

Gabriele Kotsis
President
Association for Computing Machinery



Association for
Computing Machinery

Advancing Computing as a Science & Profession

SHAPE THE FUTURE OF COMPUTING. JOIN ACM TODAY.

www.acm.org/join/CAPP

SELECT ONE MEMBERSHIP OPTION

ACM PROFESSIONAL MEMBERSHIP:

- Professional Membership: \$99 USD
- Professional Membership plus ACM Digital Library: \$198 USD (\$99 dues + \$99 DL)

ACM STUDENT MEMBERSHIP:

- Student Membership: \$19 USD
- Student Membership plus ACM Digital Library: \$42 USD
- Student Membership plus Print *CACM* Magazine: \$42 USD
- Student Membership with ACM Digital Library plus Print *CACM* Magazine: \$62 USD

- Join ACM-W:** ACM-W supports, celebrates, and advocates internationally for the full engagement of women in computing. Membership in ACM-W is open to all ACM members and is free of charge.

PAYMENT INFORMATION

Name

Mailing Address

City/State/Province

ZIP/Postal Code/Country

- Please do not release my postal address to third parties

Email Address

- Yes, please send me ACM Announcements via email
- No, please do not send me ACM Announcements via email

- AMEX VISA/MasterCard Check/money order

Credit Card #

Exp. Date

Signature

Purposes of ACM

ACM is dedicated to:

- 1) Advancing the art, science, engineering, and application of information technology
- 2) Fostering the open interchange of information to serve both professionals and the public
- 3) Promoting the highest professional and ethics standards

By joining ACM, I agree to abide by ACM's Code of Ethics (www.acm.org/code-of-ethics) and ACM's Policy Against Harassment (www.acm.org/about-acm/policy-against-harassment).

I acknowledge ACM's Policy Against Harassment and agree that behavior such as the following will constitute grounds for actions against me:

- Abusive action directed at an individual, such as threats, intimidation, or bullying
- Racism, homophobia, or other behavior that discriminates against a group or class of people
- Sexual harassment of any kind, such as unwelcome sexual advances or words/actions of a sexual nature

BE CREATIVE. STAY CONNECTED. KEEP INVENTING.



ACM General Post Office
P.O. Box 30777
New York, NY 10087-0777

1-800-342-6626 (US & Canada)
1-212-626-0500 (Global)
Hours: 8:30AM - 4:30PM (US EST)

Fax: 212-944-1318
acmhelp@acm.org
www.acm.org/join/CAPP

Technology Strategy and Management

From Remote Work to Working From Anywhere

Tracing temporary work modifications resulting in permanent organizational changes.

THE COVID-19 PANDEMIC made remote working a sudden necessity for many employers and employees in 2020. This shock resulted in 35%–50% of all U.S. employees working entirely or partly from home by May 2020.^{1,2,4} Information technology played a central role, with Internet connection at home making the transition to remote work remarkably unproblematic for most people. But many surveys carried out to enquire about how individuals view remote work demonstrate that its impact can be a double-edged sword. Some are loving it, with flexible schedules, no long commute, and more time with family. But others are unhappy with loneliness and the blurred boundary between work and leisure.^{5,6} How can we make sense of these mixed pressures? What are the different factors that have affected and will continue to influence the way we work? This column considers what remote working has meant before the pandemic, and its likely transformation in a

post-pandemic world. I argue that the institutionalization of “remote work” as “working from anywhere” will require deep changes in organizational life.

Trend Toward Remote Working Predates the COVID-19 Shock

Remote working is not a new idea. The COVID-19 pandemic accelerated changes already under way, and pushed things over a “tipping point.” However, challenges remain for workers, including those in the IT and software industry, not least because of the sudden and unplanned way in which this happened, alongside the furloughing of jobs and the closure of schools.

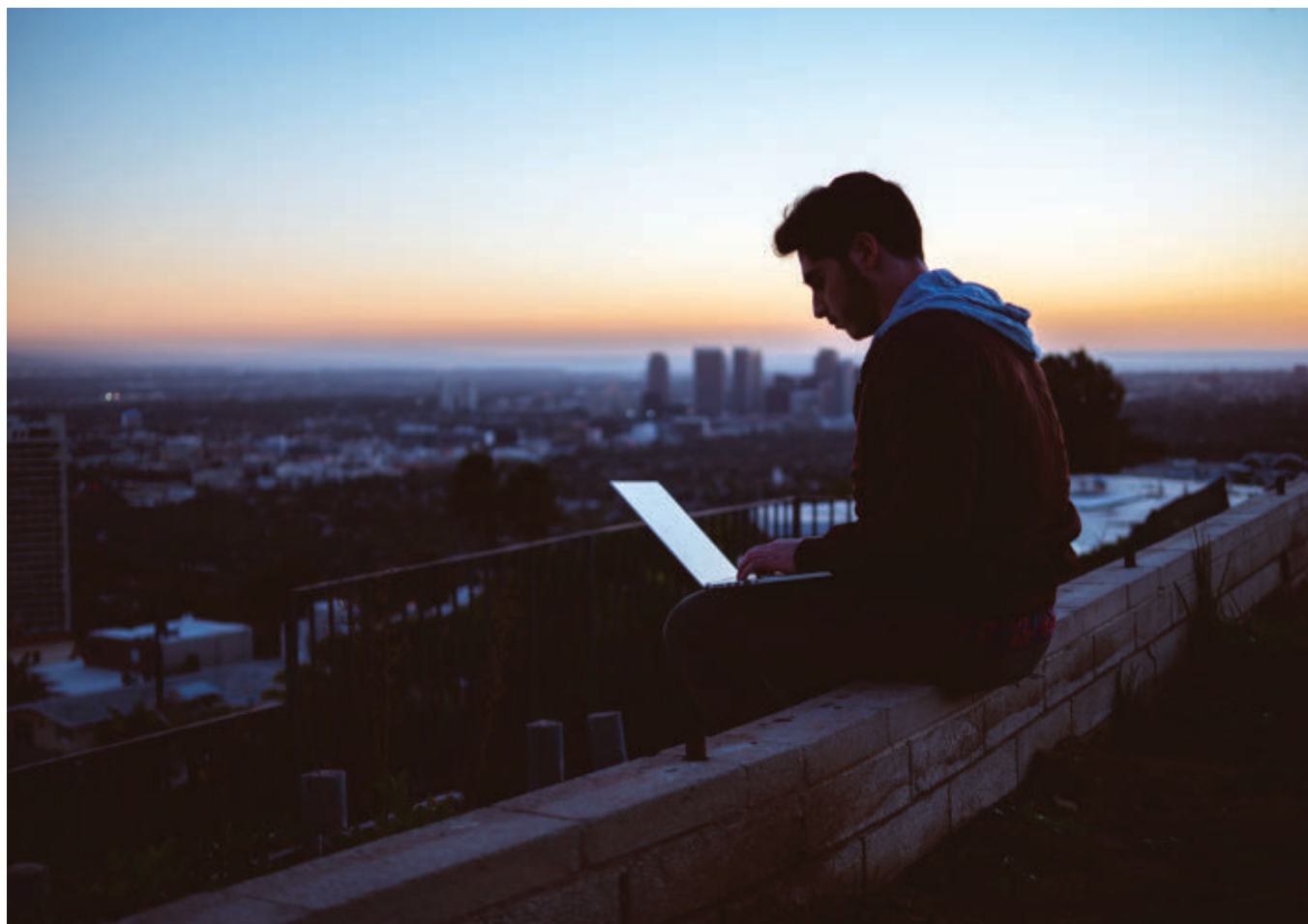
The relationship between technology and work has been subject to change for a long time. During the early phases of the first industrial revolution, we saw the “putting out” of work to homeworkers. With the rise of large manufacturing plants and offices in the 20th century, many employees experienced a clear separation between workplace and

home. In this context, remote working came to be defined by the Cambridge English Dictionary as “the practice of an employee working at their home, or in some other place that is not an organization’s usual place of business.”

By the 1980s, remote office work—work-from-home (WFH)—was considered an extension of flexible work arrangements⁹ alongside part-time work, which enabled workers, especially female workers, to balance work and childcare. With the Internet enabling connectivity since the 1990s, remote work morphed into offshoring to low-cost global locations. Offshored work included office work at call centers and software engineering centers, but also freelancing in design, data entry, programming, and translation using platforms such as Upwork, LinkedIn ProFinder, and Fiverr.

Remote Work Varies by Type of Work

Of course, remote working patterns vary by type of work. As early as the



1980s, when home-based digital technology was limited, Margrethe Olson published in this magazine a careful study of remote office work.⁹ She documented the characteristics of jobs that predisposed job holders to work remotely. They included minimal physical requirements of the job in the form of a telephone and a terminal, high degree of individual control over work, well-defined milestones and deliverables, the need for concentration, and low need for communication. All of these characteristics still apply to remote work and freelancing today.

More recently, there is robust evidence of how remote work varies by occupation. According to the American Time Use Survey (ATUS) in the U.S., for example, managerial, professional, and related (MPR) occupations—a broad category including managers, financial analysts, engineers, computer programmers, and lawyers—were found to be among those whose jobs could be performed in a variety of locations, including at home.^{7–8} In 2013–2017, workers in these occupations accounted for 41% of

the U.S. workforce. Among them, people employed in healthcare (physicians, nurses, therapists, for example) and technical (such as architectural, engineering) occupations were among the most likely to work only at their workplaces. By contrast, people employed in arts, design, entertainment, sports, and media and in education, training, and library occupations were the least likely to work only at their workplace on days they spent time working.

Remote working trends predate the pandemic. Specifically, the share of MPR workers who worked solely at their workplace declined from 46% in 2003–2007 to 41% in 2013–2017. Over the same period, the proportion of MPR workers who worked only at home on a given day increased from 10% to 13%. The ATUS results for 2020 are not yet published, but evidence from the Real-Time Population Survey provides expected outcomes. Comparing February and May 2020, the proportion of employees who worked at home every day increased from 5% to 24% in healthcare, whereas the home-every-day pro-

portions increased from 13% to 60% in professional and business services, and from 11% to 61% in the financial and insurance sector.²

Remote Work Facilitated by Digital Technology

Digital technology has undoubtedly made it increasingly feasible for companies to hire workers remotely to get tasks done. Computers are faster and cheaper, and stable broadband Internet is widely available in most locations. Moreover, advances in video chat platforms, cloud-based services, and desktop virtualization have facilitated remote collaboration in a variety of knowledge work including R&D, product development, and marketing. Gone are the days when we used to gawk at how “real” the video conferencing rooms were, with “real” meaning visual 3D and no delay in audiovisual transmission.

These technological advances account for how the smooth transition to remote working was in spring 2020 for the majority of workers. Companies

that were not prepared accelerated investments in cloud-based software tools, eSignatures for documents, and purchase of equipment to support working from home. With a surge in demand for these specific technology areas, start-ups in cloud-based technologies attracted early-stage investment funding, and represented a bright spot in a much gloomier investment climate, according to a 2020 survey by 500 Start-ups, a global venture capital fund and seed accelerator.^a Moreover, COVID-19 has shifted patent applications toward technologies that support working from home.⁴ This bodes well for enabling workers to communicate better with each other.

Tasks, Not Jobs, Will Be Subject to Remote Working

A phenomenon related to remote work is freelancing. According to a survey commissioned by Upwork and Freelancers Union, 57 million people, representing 35% of the U.S. workforce, engaged in freelancing in 2019.^b They generated \$1 trillion in income, or 5% of the U.S. economy, in 2019. The share of full-time freelancers increased from 17% in 2014 to 28%. The most common types of freelance work are in skilled services, with 45% of freelancers providing programming, marketing, and consulting services. Freelancers, while attracted to freelancing for the freedom to do work from anywhere, face a variety of location choices: 27% of skilled freelancers do all of their work remotely, while 19% do none or little of their work remotely. Remote means locations away from traditional offices, including home but also coffee shops and co-working spaces.

One way of understanding what freelancers do is to disaggregate a job into tasks that have the characteristics of minimal physical requirement, full control, well-defined milestones and deliverables, and so forth that Margrethe Olson identified in the 1980s. For example, a programmer may interact with clients to understand their requirements; she then works on her own to develop the codes, before making further refinements in accordance with customer feedback. Separating out the interactive

a 500Startup. (2020). The impact of COVID-19 on the early-stage investment climate.

b Freelancing in America; <https://bit.ly/37jzxad>

There is strong evidence that most of us wish to retain an element of remote working in our jobs.

tasks and work-alone tasks is possible, and it is the latter that are more easily subjected to remote working. Thus, while working some days at home may be a matter of lifestyle choice, there are also task-based rationale for a part-home-part-office work mode.

Non-Technological Norms Will Be Important for the Long-Term

While certain task characteristics and digital technology facilitate or undermine remote work, ultimately what will determine the “new normal” in remote working will be our views and social norms about work modes. For everyone, there will be importance attached to social interaction in human organizations. We draw boundaries so that one either belongs or does not belong to an organization. And however good the technology might become, we will miss casual encounters in office corridors and the proverbial “water cooler” conversations.

Toward Working from Anywhere

There is strong evidence that most of us wish to retain an element of remote working in our jobs. As many as 98% of those surveyed in a study want to have the option to work remotely for the rest of their lives.^c Another survey found that by 2022, employers on average were planning for employees to spend about one day per week from home, while the average worker would like to work from home approximately two full days per week.¹

The COVID-19 pandemic created a discrete shock to our work patterns, with people forced to work from home

c The 2020 state of remote work; <https://bit.ly/37lsL3B>

in an unplanned manner. However, trends toward remote work, particularly for certain occupations, predates the pandemic, and are characterized by the existence of certain tasks that are easy to carry out remotely. It becomes clear that in most jobs, there are tasks that can be carried out remotely—manipulating data, writing codes and reports and so forth—and tasks that are better carried out in social spaces with other co-workers, such as brainstorming and performance that require feedback. If we derive our well-being from balancing these two types of tasks, then remote work will be part of a hybrid model combining remote and in-office working.

But apart from the technical requirements of task execution, social norms will have to change, with less stigma attached to working from home, with a history of association with shirking or hiding or not being a good citizen.¹ Gender roles will have to be re-examined, as today it is still women with small children who would wish to get out of the home and into the office. Digital technology of the future, with cloud computing, has much to facilitate such hybrid working. But it will take a lot more than giving employees laptops and broadband connections for “working from anywhere” to take hold. **□**

References

- Barrero, J.M., Bloom, N., and Davis, S.J. Why working from home will stick. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2020-174), (2020).
- Bick, A., Blandin, A., and Mertens, K. Work from home after the Covid-19 outbreak. *The Federal Reserve Bank of Dallas, Research Paper 2017*, (2020).
- Bloom, N., Davis, S.J., and Zhestkova, Y. COVID-19 shifted patent applications toward technologies that support working from home. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2020-133), (2020).
- Brynjolfsson, E. et al. Covid-19 and remote work: An early look at U.S. data. *NBER Working Paper No. 27344*; (June 2020).
- IBM. Covid-19 and the future of business *Insights series*: IBM Institute for Business Value; The 2020 state of remote work; <https://bit.ly/2NboGbw>
- IBM. What 12,000 employees have to say about the future of remote work; <https://on.bcg.com/2ZkinEP>
- Krantz-Kent, R.M. Where did workers perform their jobs in the early 21st century? *Monthly Labor Review*, 1-10 (2019).
- Krantz-Kent, R.M. Where people worked, 2003 to 2007. *U.S. Bureau of Labor Statistics*. Washington, D.C. (2009).
- Olson, M.H. Remote office work: Changing work patterns in space and time. *Commun. ACM* 26, 3 (Mar. 1983), 182–187.

Mari Sako (mari.sako@sbs.ox.ac.uk) is Professor of Management Studies at Saïd Business School, University of Oxford, U.K.

Copyright held by author.

Broadening Participation

Reflections on Black in Computing

Seeking to improve systemic fairness in the computing realm.

IN JUNE 2020, a community of Black people in computing from around the world published an open letter,^a initiated by the authors, and a call for action^b to the global computing community. The letter began with, “The recent killing of George Floyd by Minneapolis Police has sparked a movement that began at the birth of our nation. Though George Floyd may have been the most recent instance, we should not forget the lives of Breonna Taylor, Ahmaud Arbery, Nina Pop, Tony McDade, Sandra Bland, Trayvon Martin, Aiyana Stanley-Jones, Philando Castille, Tanisha Anderson, Atatiana Jefferson, Eric Garner, Charleena Lyles, Eula Love, Michael Brown, Khalif Browder, Botham Jean, Tamir Rice, Latasha Harlins, Amadou Diallo, Mary Turner, Emmett Till, and too many other Black people who have been murdered ...”

At the time, we reflected on this history of the killing of Black people in the U.S. and noted that these killings not only show the ultimate outcomes and harms that racist systems and institutions have on Black people, but they also spotlight the constant emotional and psychological strain that Black Americans endure. The accumulated experience of the Black computer science community highlights the magnitude of injustices that countless members of our community experience. During the course of performing our



jobs, we endure general mistreatment and we face a lack of support, demonization, and erasure of our (Black) academic and professional expertise. We know it is important that we persist in raising concerns about discrimination and prejudices that Black professionals experience, which are often common practice in the field. Further, we are acutely aware that organizational policies are currently optimized to exclude non-white males.

After our call to action, more than 700 signatures from individuals representing the breadth and depth of the computing and technology communities were received. People from academia, industry, government, and non-profit sectors signed, in solidarity, with the sentiment that we must do more. Accompanying the letter and call to ac-

tion was a definitive list of actionable steps individuals and organizations can take to redress systemic racism that exists in our profession, and beyond.

Months Later

As we all grappled with the compounding and collective grief of the pandemic and institutional harm done to Black people in the U.S. and in other majority white countries, there were a plethora of statements made in support of Black employees, students, business owners, and founders, as well as the broader Black Lives Matter movement. Months later, as we reflect on those statements (and their promises), we are curious about the action, the follow-up, and the changes in policy and practice that will institutionalize the commitments, catalyze the change needed, support Black

a See <https://blackincomputing.org>

b See <https://blackincomputing.org/action-item-list>

lives, and create an environment that is equitable and fair for all. To date, there has been little real action beyond initial statements.

As students, teachers, mathematicians, scientists, technologists, and engineers, we learn there is no need for “culture” in our field. Ones and zeros, the scientific method, and meritocracy form the basis of our discipline. Computing is “neutral.” We know this is not true, which means that computing, as an institution, is still a long way from realizing its promise to make the world a better place. We also know that our field does not exist in a vacuum. The structural and institutional racism that has brought the nation to this point is also rooted in our discipline. We see AI and big data used to target the historically disadvantaged. The technologies we help create to benefit society are also disrupting Black communities through the embodiment of systemic bias, prejudice, and the proliferation of racial profiling. We see machine learning algorithms—rather, those who are developing the algorithms—routinely identify Black people as animals and criminals. Technology that we develop is used to further intergenerational inequality by systematizing segregation in housing, lending, admissions, policymaking, healthcare, and hiring practices.

We know better. We are not fooled by the doublespeak, the pleas of ignorance, and the excuses for the technological systems that are deployed into the world. We know that the advances in computing are transforming the way we all live, work, and learn. We also know that we cannot ask for equal opportunity for anyone without demanding equal opportunity for everyone. We know that in the same way computing can be used to stack the deck against Black people, it can also be used to stack the deck against anyone.

Call To Action

Today, we are issuing another call to action to the individuals, organizations, educational institutions, and companies in the computing ecosystem to address the systemic and structural inequities that Black people experience. In issuing this call we ask the community to:

- Create unbiased and welcoming learning and work environments that

We know that in the same way computing can be used to stack the deck against Black people, it can also be used to stack the deck against anyone.

allow Black people to be their authentic and whole selves, learning and working without experiencing racism and bias.

- Commit to addressing the systemic and institutional racism that has led Black people in computing to be pushed out of the field or to exit the field to pursue alternative careers.

- Address issues related to corporate, organizational, and educational culture and climate to create welcoming and comfortable spaces for Black people and prioritize the health and well-being of all computing students, employees, faculty, volunteers, and entrepreneurs.

There is a role for each of us to build stronger, more creative, and more inclusive communities, as described here.

Individuals can acknowledge the presence of Black colleagues, be open to new ideas and perspectives, and be their advocate or ally in times of discrimination and otherwise. Everyone can reflect on privileges they might have such that we can move toward eliminating double standards.

Educational Institutions can ensure that perpetrators of a toxic environment face consequences for their actions and that the injured parties are supported, not blamed, ostracized, and forced out. They also can reset their procedures and systems to be equitable and just, ensuring institutional power does not enable the subjective mistreatment of Black students, employees, postdocs, and faculty. They can integrate an equitable, fair, and just racial lens to every major milestone along the academic path to ensure that bias, prejudice, and discrimi-

nation do not play a part in anyone’s journey.

Organizations that receive public funding can ensure they are providing equal opportunity in compliance with existing civil rights statutes, including but not limited to the Civil Rights Act of 1964, the Education Amendments Act of 1972, and the Americans with Disabilities Act of 1990. They must also go beyond compliance and lip service to implementing systems and policies that realize actual outcomes that demonstrate progress on attracting, supporting, keeping, and promoting Blacks.

Corporations can start taking meaningful actions toward solving the racism problem that permeates their culture, leadership, staff, and tools. Publishing diversity metrics and issuing statements of performative progressiveness have not yielded progress or improved the lives of Black employees and entrepreneurs. Corporations need to change, positively, and/or eliminate policies and procedures that are weaponized against Blacks. They also need to consistently and fairly hold those that cause harm accountable.

Communities can establish equal opportunity review structures that are responsible for collecting and analyzing data to certify equitable outcomes by institutions, companies, and organizations in computing. These communities must also offer support and be strong voices for change and agents of actions for those who are harmed.

As we did in June 2020, we ask that you translate the public statements^c into public action to support the Black professional communities toward achieving systemic fairness in computing. ■

c See <https://bit.ly/3s3nDsV>

Quincy Brown (quincykbrown@gmail.com) is the co-founder of blackcomputeHER.org in Upper Marlboro, MD, USA.

Tyrone Grandison (tgrandison@data-driven.institute) is the founder of The Data-Driven Institute and co-founder of The Human Collaborative in Seattle, WA, USA.

Jamika D. Burge (jamika.burge@gmail.com) is the co-founder of blackcomputeHER.org and the founder and principal of Design & Technology Concepts in Alexandria, VA, USA.

Odest Chadwicke Jenkins (ocj@umich.edu) is a Professor in the Department of Computer Science and Engineering at the University of Michigan, Ann Arbor, MI, USA.

Tawanna Dillahunt (tdillahunt@umich.edu) is an Associate Professor at the University of Michigan School of Information, Ann Arbor, MI, USA.



Kode Vicious

The Non-Psychopath's Guide to Managing an Open Source Project

Respect your staff, learn from others, and know when to let go.

Dear KV,

In the past you have written about managers and management in your columns, and I note that, on balance, you do not often have nice things to say. Even when you write about management in noncorporate culture—such as open source projects—you describe it in such a way that no technically minded person would ever want to get involved in managing people or projects.

I am writing now because I have recently been convinced (I fear you will say duped) to accept the leadership position in a long-running, open source project—and not only that, I am working with both volunteers and paid staff. I agreed to this for what are probably the usual reasons, but most of all, out of a desire to help the project continue to grow and do well. Naive, perhaps, but that was the reason at the time I agreed to the job.

Now I find myself deluged on all sides with technical problems—some big and some small—as well as people problems among both the volunteers and the paid folks I work with. I feel like everything is out of my control and an emergency all at the same time, although, of course, I say, “Everything is fine,” as I don’t want to let on that it’s not. While I am not at the end of my wits, I fear that after a year in this position, I definitely will be, which I think is the reason the person in this position before me stepped down after two years.



I have tried watching videos that might help me, including “How Open Source Projects Survive Poisonous People,” the Google TechTalk by Ben Collins-Sussman and Brian Fitzpatrick that was mentioned in one of your columns (*acmqueue*, January-February 2008), and I have skimmed a few books on managing, but the books really are not engaging at all and seem to make my head hurt even worse.

Is there some knack to managing

open source projects that I can acquire without feeling like I am running the gauntlet?

Duped and Deluged

Dear Duped,

Transitioning from one of the technical faithful to one of the hated PHBs (pointy-haired bosses), whether in the corporate or the open source world, is truly a difficult transition. Unless you

acm

Advertise with ACM!

Reach the innovators and thought leaders working at the cutting edge of computing and information technology through ACM's magazines, websites and newsletters.



Request a media kit with specifications and pricing:

Ilia Rodriguez
+1 212-626-0686
acmm mediasales@acm.org

acm

media

are a type who has always been meant for the C-suite—where, it is reported, a certain percentage of CEOs are psychopaths—it is going to take a lot of work and a lot of patience, mostly with yourself, to make this transition. Doing something “for the good of (blank)” usually means you are sublimating your own needs to the needs of others, and if you don’t acknowledge that, you are going to get smacked and surprised by your own reactions to people very, very quickly.

As for the fact that I “do not often have nice things to say” about managers and management, well, go back and read a few more of my pieces, and please let me know just how many of them actually say *anything* nice. Go ahead, I will wait; it will not take long.

Probably the first hurdle most technical people have to pass in helping others get their work done—which is what good managers actually do—is forgetting all the lessons they learned from typical managers they worked with in the past, and learning some new ones.

For some reason, technology is full of managers who can be classified only as, well, awful. KV thinks this is because most technologists like interacting with technology more than they like interacting with humans. When my program crashes for the 30th time with some Heisenbug, and I scream a litany of curse words that would reduce a Navy salt to tears, no one, other than perhaps those within earshot who didn’t really know that those positions were possible, is really hurt. If you were to direct even one-tenth of that anger at a typical human being, you would both wound them emotionally and make it harder to work with them in the future.

Lesson number one, people are not machines. Now, I know some smart-aleck (normally that would be KV) will point out that we are all just biological machines, but until someone produces a credible architecture manual and instruction set for human emotions, let’s just say we need to treat people differently from computers. I mean, you don’t put your latest Hacker Space L33T sticker on the person in the next cubicle as you would on your laptop, right? Right.

Given what we know of typical management in technical circles, how then do you find good examples to work

from? Humans learn by copying other humans, whether it is how to code, or how to treat others, or how to organize work. You have to decide, first, what success in an endeavor means to you. If success is having the most money, then become a psychopath (the *DSM-V* has the definition) and climb the corporate ladder. If success is managing a group of people who are generally happy working together on a project, then I suggest looking for guidance from successful volunteer projects outside of technology.

One of the best managers I ever met managed volunteer theater before going into technology and running a large training group. During the five years we worked together, his group was consistently the happiest and most productive. Not that they were a bunch of Pollyannas wandering the halls in a state of bliss, but his team did have the lowest turnover and the lowest absentee rate in the company. When I asked about this, he explained that once you have worked with emotionally overwrought prima donnas who are doing what they love for zero money, and probably not much more fame, you can manage anyone. This boils down to knowing what each person is capable of, what engages them most, and then providing them as much work based on those criteria as they can handle.

The role of the manager is not to tell people what to do, but to give them the opportunity to do the work they excel at and that serves the goal (a.k.a. the project) as much as possible. You will never be able to make this happen 100% of the time, but it should always be your goal.

It is also not the case that there are no good examples to follow, though there may be fewer than we might like. I particularly like the leadership style of Guido van Rossum, who developed and then led the Python project for nearly 30 years. Watching the project from the outside, I believe it managed to maintain an open and collaborative structure throughout its history, and, in what I always consider the sign of a good leader, Guido stepped down from the leadership position after transitioning those functions onto others on the project. Knowing when to let go is another lesson successful managers must learn.

One of the great things about open source's open nature is that—via the mailing-list archives—you can see how decisions were made and how issues got resolved. Look to other projects that have high cohesion and good culture, and you will probably find a useful example or two. Seek out those who were led by diamonds in the rough and see how they made the same transition you are making.

And now to two things that everybody lies about. The first is that taking on a leadership position gives you more control. Actually, taking on a leadership position means dropping some control rather than gaining it. Technologists overall have some serious control issues, and while this may be a broad generalization, I recommend you stop right here and think about how you feel when something you are working on is out of your control. Heart rate up? Anger level up? Sweaty palms?

Leaders who try to maintain complete control over every aspect of a project most often fail. Building the ability to trust those around you and delegate to them is something you'd better do right now or that drowning feeling you describe in your letter is going to get worse. The harder you hold on to control, the less you will have until you either quit or are pushed out by those around you who will see you as more of a menace than a help. So, another lesson is learning how to delegate: As a leader, you must be able to take the long view and trust others with day-to-day minutiae. I like to call this "Seeing the Greater Narrative," but I am sure it has a more boring name somewhere.

The second lie is, "You can continue doing the same technical work you always did." This is, in fact, the biggest lie told to people when they first transition from a purely technological role into a managerial one. It is complete and utter nonsense, and KV may or may not have actually said that (or perhaps worse) when this little gem was tried on him many years ago. Unless you were not doing much when you were doing technical work, adding leadership and management responsibilities, with all the learning that entails, is initially going to crush your technological productivity. So, lesson four: As you progress in the leadership position, it will be important to stay technically adept, even

Being the leader of an open source project is not something you are ever going to learn in school, but it can be learned.

if you are no longer dumping KLOCs into a system.

This, again, is where building trust with those around you helps. Let them explain to you the things you do not understand but want to, either to do your job or because as a naturally technically minded person this keeps you engaged with the project goals. Too often I have seen people go into management and become truly dulled to the finer points of technology, and then they really do become pointy-haired bosses.

Being the leader of an open source project is not something you are ever going to learn in school, but it can be learned, and it can be done well—if only more people would assimilate these four lessons: People are not machines; delegate; know when to give up control; and stay technically engaged.

Or you can become a psychopath. I hear CEOs make serious bank.

KV

Related articles on queue.acm.org

Forked Over

Kode Vicious

<https://queue.acm.org/detail.cfm?id=2611431>

The Age of Corporate Open Source Enlightenment

Paul Ferris

<https://queue.acm.org/detail.cfm?id=945124>

A Chance Gardener

Kode Vicious

<https://queue.acm.org/detail.cfm?id=3286730>

George V. Neville-Neil (kv@acm.org) is the proprietor of Neville-Neil Consulting and co-chair of the ACM *Queue* editorial board. He works on networking and operating systems code for fun and profit, teaches courses on various programming-related subjects, and encourages your comments, quips, and code snips pertaining to his *Communications* column.

Copyright held by author.



Digital Threats: Research and Practice

Digital Threats: Research and Practice (DTRAP) is a peer-reviewed journal that targets the prevention, identification, mitigation, and elimination of digital threats. DTRAP aims to bridge the gap between academic research and industry practice. Accordingly, the journal welcomes manuscripts that address extant digital threats, rather than laboratory models of potential threats, and presents reproducible results pertaining to real-world threats.



For further information and to submit your manuscript, visit dtrap.acm.org

Historical Reflections When Hackers Were Heroes

The complex legacy of Steven Levy's obsessive programmers.

FORTY YEARS AGO, the word “hacker” was little known. Its march from obscurity to newspaper headlines owes a great deal to tech journalist Steven Levy, who in 1984 defied the advice of his publisher to call his first book *Hackers: Heroes of the Computer Revolution*.¹¹ Hackers were a subculture of computer enthusiasts for whom programming was a vocation and playing around with computers constituted a lifestyle. Levy locates the origins of hacker culture among MIT undergraduates of the late-1950s and 1960s, before tracing its development through the Californian personal computer movement of the 1970s and the home videogame industry of the early 1980s. (The most common current meaning of hacker, online thieves and vandals, was not established until a few years later).

Hackers was published only three years after Tracy Kidder’s *The Soul of a New Machine*, explored in my last column (January 2021, p. 32–37), but a lot had changed during the interval. Kidder’s assumed readers had never seen a minicomputer, still less designed one. By 1984, in contrast, the computer geek was a prominent part of popular culture. Unlike Kidder, Levy had to make people reconsider what they thought they already knew. Computers were suddenly everywhere, but they remained unfamiliar enough to inspire a host of popular books to ponder the personal and social transformations triggered by the microchip. The short-



lived home computer boom had brought computer programming into the living rooms and basements of millions of middle-class Americans, sparking warnings about the perils of computer addiction. A satirical guide, published the same year, warned of “micromania.”¹⁵ The year before, the film *Wargames* suggested computer-obsessed youth might accidentally trigger nuclear war.

Making hackers into heroes, rather than figures of fun or threat, was a bold

move. Even within the computing community, “hacker” was an insult as often as a point of pride. Managers lamented the odd work habits and unmaintainable code produced by those who programmed for love rather than money, while computer scientists decried the focus of the hackers on practice over theory. According to Levy, though, “beneath their often unimposing exteriors, they were adventurers, risk-takers, artists.” His book established a glorious lineage for the amateur programmers

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

of the 1980s, crediting their tribe with the invention of videogames, personal computing, and word processing.

The Hacker Ethic

There is a good chance you have read Levy's book, which sits perennially near the top of Amazon's computer history bestseller list and received more than 130 new citations in 2020. Even if you have not, you have probably come across his list distilling the "hacker ethic" into six bullet points. Levy proclaimed it "their gift to us" with "value even to those of us with no interest at all in computers." It goes:

- ▶ Access to computers—and anything that might teach you something about the way the world works—should be unlimited and total. Always yield to the Hands-On Imperative!

- ▶ All information should be free.

- ▶ Mistrust authority—promote decentralization.

- ▶ Hackers should be judged by their hacking, not criteria such as degrees, age, race, sex, or position.

- ▶ You can create art and beauty on a computer.

- ▶ Computers can change your life for the better.

Like an anthropologist visiting a remote tribe, Levy had the outsider perspective needed to recognize and document the core assumptions of an unfamiliar culture. In a sense those bullet points are timeless. If you are reading *Communications* you surely recognize some of these beliefs in yourself or in people you have worked or studied with. Yet reading the list is no alternative to understanding these beliefs in their original context. In this column I provide that context.

The MIT Hackers of the 1960s

The first part of the book, explaining the origins of the hacker ethic at MIT, is by far the most influential. The word *hack* was engrained in MIT's culture. Like other students, MIT students liked to play pranks. For gifted and highly competitive engineers, pranking was a chance to show off with breathtaking but pointless feats of engineering creativity. The quintessential MIT hack took place in 1984, when what appeared to be a campus police car appeared on top of the Great Dome.

Computer programs that demon-

What was exceptional about MIT was not that it had a computer or that unkempt programmers were devising impressive tricks. It was that MIT had enough computers that a couple of surplus machines could be left out for members of the community to play with.

strated great skill yet served no apparent utilitarian function could also be "hacks." The computer hacker community formed around MIT's TX-0 computer from about 1958, expanding when one of the first minicomputers, a Digital Equipment Corporation PDP-1, was installed in 1962. MIT was then a world leader in computing, thanks to clusters of expertise built around Project Whirlwind (an early digital computer that became a prototype of the SAGE air defense network), pioneering work on timesharing (which became Project MAC), and the Artificial Intelligence lab founded by John McCarthy and Marvin Minsky.

The shared joy of the hackers was manipulating the functioning of formal, rule-based systems to produce unanticipated results. Levy describes them hacking *Robert's Rules of Order* during meetings of the MIT Tech Model Railroad Club, hacking the English language to produce words that should logically exist like *winnitude*; even hacking the Chinese symbols printed on the menu of their favorite restaurant to create the inedible "sweet and sour bitter melon."

Computer programming offered unparalleled opportunities for the virtuoso manipulation of symbols. Levy demonstrates this by showing the hackers shaving instructions from a decimal print routine. A communal frenzy of competitive programming over several weeks culminated in the quietly triumphant posting of an optimal routine on a noticeboard. This was not so unusual: getting a computer to do anything during the 1950s required feats of efficient programming. When interviewed, computing pioneers often recall the joy of loader programs squeezed onto a single punched card, subroutine calling mechanisms that saved a few instructions, or assemblers that automatically distributed instructions around a magnetic drum so that they would be read just in time to be executed. Many celebrated computer scientists began as systems programmers, a group known for its unconventional appearance. As early as 1958, before either hackers or hippies were documented, a *Business Week* article complained that "computers have been in the wrong hands. Operations were left to the long-hairs—electronics engineers and mathematicians ..."¹ In 1966, as the Data Processing Management Association, a group for supervisors of administrative computing centers, began to contemplate a relationship with ACM one of its leaders repeated a rumor that ACM members "were part of the sweat-shirt and sneaker group."⁴

What was exceptional about MIT was not that it had a computer or that unkempt programmers were devising impressive tricks. It was that MIT had enough computers that a couple of surplus machines could be left out for members of the community to play with. Most computers were in the hands of specialist operators, who in an oddly anti-Catholic metaphor, Levy dismisses as a "priesthood" standing between the faithful and direct access to instruments of salvation. Students would submit programs and get results, but never touch the computer or interact with it directly. Even universities that treated computers like lab equipment, letting researchers sign up for an hour or two with the machine, required a documented purpose.

Levy carefully distinguishes his hackers from the "officially sanctioned

users” for whom computing was a means to solve research problems and publish papers rather than an end in itself. For the hackers, most of whom started as undergraduates, computer programming was not an aid to formal studies but an alternative. Many of Levy’s characters drop out of college to spend more time playing with the machines. One hacker implements the first LISP interpreter, others leave for California to develop operating systems at Berkeley, and Minsky hires several to produce software for his lab. Yet Levy keeps their official work on timesharing and AI offstage, instead focusing on their nighttime pursuits. They are a rich cast, and Levy does a wonderful job of bringing them to life as quirky individuals with their own characteristics rather than as interchangeable geeks.

Freed of the need to demonstrate any useful purpose for their programs the hackers pioneered applications of computer technology that became widespread once hardware costs dropped. It helped that the PDP-1 was equipped with a vector-based graphical display, an unusual capability for the era. One of their programs was the “expensive typewriter,” which used the screen to edit program code. It made little economic sense to tie up an entire computer to program, which is why the usual approach was to write code out with pencils, to be punched onto cards or paper tape. Another was the “expensive calculator” that replicated the interactive functioning of an electromechanical desk calculator on a device hundreds of times more costly.

Levy spends an entire chapter on the most glorious misapplication of resources undertaken by the hacker collective: the video game *Spacewar* (or, as its main author Steve Russell likes to call it, “Spacewar!”) Inspired by old science fiction books, the hackers programmed routines to simulate and visualize the movement of rocket ships in space. Adding photon torpedoes, an intricate starfield background, and the gravitational pull of a star created an addictive combat game. This was not quite the first videogame, but it was the first to matter. DEC began to distribute the code as a diagnostic for its computers, spreading it to many other sites. *Spacewar* was profiled in a 1972 [Rolling Stone](#) article by Stewart Brand,

Freed of the need to demonstrate any useful purpose for their programs the hackers pioneered applications of computer technology that became widespread once hardware costs dropped.

founder of the *Whole Earth Catalog*, bringing it more fame.² By then, Nolan Bushnell and Ted Dabney had reimplemented the game in hardware as *Computer Space*, the first coin-operated video arcade game. It proved too complicated for drunken users in bars, but Nolan and Dabney kickstarted the video arcade industry with their next release: *Pong*.

Celebrating Hacker Culture

The original hackers were neither destructive nor dedicated to the pilfering of proprietary data, unlike the online vandals and criminals who later appropriated the word, but they were quite literally antisocial. Levy describes their lack of respect for any rules or conventions that might limit their access to technology or prevent them from reconfiguring systems. They are seen bypassing locked doors, reprogramming elevators, and appropriating tools.

Most technology writers can be pegged as critics or cheerleaders. To the cheerleaders, new technologies open utopian possibilities and unlock human potential. To the critics, each new technology is a study in unintended consequences or a way to reinforce injustice and oppression. Levy is not uncritical, but he is unmistakably more interested in capturing how his protag-

onists view the world than in hectoring them. The book’s subtitle, “Heroes of the Computer Revolution,” does not admit very much nuance.

Not all observers of hacker culture were so accepting. Levy rejected MIT professor Joseph Weizenbaum’s portrayal of the institute’s “computer bums” (a term borrowed from Brand), which recalled the sordid opium dens found in Victorian novels: “bright, young men of disheveled appearance, often with sunken glowing eyes, can be seen sitting at computer consoles, their arms tensed and waiting to fire their fingers, already poised to strike, at the buttons and keys on which their attention seems to be as riveted as a gambler’s on the rolling dice Their food, if they arrange it, is brought to them: coffee, Cokes, sandwiches. If possible, they sleep on cots near the computer Their rumpled clothes, their unwashed and unshaven faces, and their uncombed hair all testify that they are oblivious to their bodies and to the world in which they move.”¹⁸

MIT professor Sherry Turkle presented an equally biting picture of MIT’s hacker culture in her ethnographic study *The Second Self: Computers and the Human Spirit*, another classic study of early computer use.¹⁶ As a humanist joining MIT’s faculty she had “immersed herself in a world that was altogether strange to me.” Turkle spent most of the book exploring the cognitive possibilities computing opened for education and personal development. Yet she used the hackers primarily as a cautionary illustration of what happens when human development goes wrong.

Hacker life was for the most part celibate, but it was nevertheless highly gendered. Levy writes that “computing was much more *important* than getting involved in a romantic relationship. It was a question of priorities. Hacking had replaced sex in their lives.” Women were almost invisible, as hackers “formed an exclusively male culture.” “The sad fact,” notes Levy, “was that there never was a star-quality female hacker. There were women programmers, and some of them were good, but none seemed to take hacking as a holy calling ...” Levy’s silence on other matters communicates that his hackers were white and that the sex they were not having was with women, the default

assumptions of that era if not, thankfully, of ours.

Hackers had created a new masculine space as culturally distinctive as the Catholic priesthood or the U.S. Marine Corps. Turkle judges the hackers more harshly than Levy for these choices. Even within the dysfunctional culture of MIT, she suggests, computer science students were the “ostracized ... archetypal nerds, loners, and losers.”¹⁵ Her chapter “Hackers: Loving the Machine for Itself” begins by describing MIT’s anti-beauty pageant, an annual competition to choose “the ugliest man on campus.” This, she suggests, is evidence of a social illness of self-loathing that “accepts and defensively asserts the need for a severed connection between science and sensuality.” According to Turkle, hackers had “got stuck” part way through the normal course of psychological development, in which adults make accommodations with what hackers called the “real world” of human relationships, jobs, and personal responsibilities. That means accepting uncertain outcomes and emotional risks, by giving up the adolescent need for “perfect mastery” of a controlled world of things. Hackers refused to do this, instead creating a “highly ritualized” culture to support and normalize that choice. She was appalled by their rejection of the sensual elements of art and culture, particularly their tendency to hear music only as an expression of algorithmic progression rather than an activity in which human emotion and instrument tuning were important.

Levy does acknowledge some negative aspects of hacker society. Hackers judged each other purely on programming skill and commitment to hacking, rather than on more conventional social markers. They were elitist, making harsh judgments of “winners and losers” based on an ethical code that privileges coding ability and commitment to programming over all other virtues.

The closest Levy came to direct criticism was flagging the most glaring contradiction of hacker life: its relationship to the military industrial complex. Hacker values decried both commercialism and hierarchical authority, favoring the free exchange of code and ideas between individuals.

Yet the TX-0 had been paid for by the government. Having fulfilled its military purpose, it could be diverted for student use. And, as Levy notes, all of the AI lab’s activities, “even the most zany or anarchistic manifestations of the Hacker Ethic, had been funded by the Department of Defense.” In the late 1960s Levy’s hackers, enjoying a “Golden Age of hacking” in the AI lab above Technology Square, were mystified to see protestors outside opposing the role of computing in the Vietnam War. As Levy puts it, “a very determined solipsism reigned on the 9th floor” as the hackers denied any connection between the geopolitics of the Cold War and their military-sponsored anarchist utopia, now protected by steel barricades and electronic locks. (Solipsism is the attitude that nothing outside one’s own mind is clearly real).

The Californian Hackers of the 1970s

In the second part of the book, Levy moves to California where government and military contracts had nurtured the production of semiconductors and electronic devices by companies in what, for the first time, people were starting to call Silicon Valley. Hacker culture arrived, in his telling at least, via Stanford University’s Artificial Intelligence lab with its strong ties (including the newly constructed ARPANET) with MIT. In California, hacker culture merged both with local countercultural movements and with preexisting communities of electronics hobbyists and professionals.

Hackers is not, alas, as well engineered as *The Soul of a New Machine*. Reviewing the book for the *New York Times*, Christopher Lehmann-Haupt noted that it starts “to limp halfway through—to bog down in details that are somehow less and less exciting.”¹⁸ Part of the problem is overfamiliarity. The MIT hackers are known to most of us only through Levy’s reporting, whereas the founding of the personal computer industry was well covered in two other 1984 works: *Fire in the Valley*⁶ and Michael Moritz’s definitive rendition of the early Apple story *The Little Kingdom*.¹⁴ Dozens of subsequent retellings have followed the same basic outlines, most notably Robert X. Cringely’s

scurilously entertaining *Accidental Empires*, Walter Isaacson’s exhaustive biography of Steve Jobs, and several movies and television shows.^{3,7}

Levy’s distinctive twist was his focus on the story’s countercultural strands, personified in his central character for this section: Lee Felsenstein, a gifted electronic engineer and committed member of the Berkeley counterculture. Fred Turner’s classic book *From Counterculture to Cyberculture*, focused on Stuart Brand, showed deep connections between cybernetic ideas developed in the early Cold War and elements of the Californian counterculture of the late-1960s. Their interaction did more to shape future politics, culture, and the application of online communication than to spur the development of core computing technologies.¹⁷ Felsenstein, in contrast, provided a rare direct connection between the classic Berkeley anti-war variant of the counter culture and the emerging personal computer industry of the mid-1970s. Joining a collective that had appropriated an obsolete minicomputer but lacked the skills and work ethic to do much with it, he created the “Community Memory Project,” a short-lived online community accessed via public terminals.

Felsenstein’s commitment to time-sharing was soon tempered by the realization that the microprocessors and memory chips he planned to use for cheap video terminals could also power freestanding computers. Levy gives a vivid description of the Homebrew Computer Club, an informal group hosted on the Stanford campus that introduced the technologies of personal computing to the Bay Area community of electronics hobbyists. It inspired Felsenstein to create the Sol 20 personal computer, an elegant design optimized for easy repair even after civilization collapsed. Because Felsenstein’s business partners had little knack for business the Sol was quickly eclipsed, though he reappeared a few years later as designer of the budget-priced, suitcase-sized Osborne 1 portable computer.

Levy’s romantic attachment to the “hacker ethic,” similar in a way to Tracy Kidder’s celebration of engineers who worked to find meaning rather than to make money, creates unre-

solved tensions here. Personal computers were not given away free, but then neither were minicomputers. The original hackers relied on other people's money. The invention of computers that could be purchased by individual users broadened access to hacking and freed it from military patronage. By the early 1980s recreational programming was a feasible hobby for millions of (mostly middle class) Americans rather than the exclusive preserve of tiny communities centered on places like MIT and Stanford.^a

Felsenstein is an undeniably fascinating character. Yet Levy's insistence on personal computing as the expression of an anticommmercial, university-derived hacker ethic makes it hard for him to deal with the success of Apple, and the often overlooked Radio Shack and Commodore, in selling hundreds of thousands of computers to individual buyers by the end of the 1970s. Steve Wozniak, Apple's founding engineer, gets a fine portrait that helped to establish him in the public imagination as the embodiment of the hardware hacker, more interested in impressing fellow hackers with the elegance of his circuits than in making money. Despite Wozniak's personal virtue, implies Levy, Apple soon betrayed the hacker ethic. As Lehmann-Haupt acidly observed, "it's hard to tell whether [Levy] is celebrating the arrival of an inexpensive home computer or lamenting its astonishing profitability."

The Videogame Hackers of the Early 1980s

The third part of Levy's book is the narrowest: a portrait of the relationship between a young videogame programmer, John Harris, and his personal computer software publisher. Harris's biggest accomplishment, a skilled conversion of the arcade game Frogger, did little to alter the course of history. Instead he serves as an everyman programmer, representing the new commercial opportunities for self-trained software developers.

The shy and unworldly Harris was

a Educational timesharing systems also played an important part in democratizing computing, largely overlooked by Levy but explored recently in *Joy Lisi Rankin, A People's History of Computing in the United States* (Harvard University Press, Cambridge, MA, 2018).

part of an early-1980s generation of teenage computer programmers. Home computers were marketed as programming machines, displaying a BASIC command prompt when they were plugged in and connected to a television set. The most dedicated programmers graduated to assembly language, like the original MIT hackers. Replicating the smooth animations of coin-operated arcade games on consumer hardware required code perfectly timed to manipulate the unique quirks of each machine. The best programmers, like Harris, tended to work alone and confine themselves to the hardware of a single machine, in his case the Atari 800 for which he had to figure out undocumented features of its sound and graphics chips. Harris never fully moved on to later platforms or more modern development methods, continuing to code for the long-obsolete Atari computers.^b

Their games were distributed using a business model borrowed from rock music and book publishing—"software houses" packaged, promoted, and distributed the programs, paying royalties to their authors. Levy casts Porsche driving college dropout Ken Williams, co-founder and manager of the fast-growing publisher Sierra Online, as the villain. Williams makes millions of dollars from the efforts of Harris and the other young programmers. Although Williams works hard to "get Harris laid," he resents paying a 30% royalty and hates being reliant on unpredictable hackers. Williams therefore colludes with venture capitalists, hires a professional manager to bring order, and flirts with software engineering methods taken from large corporate projects. His industry embraces unhacker like behavior such as intellectual property lawsuits and copy protection.

I think of this section as a long magazine article that was somewhat arbitrarily bound in the same volume as Levy's historical research. Lehmann-Haupt complained that each section of the book "seems more trivial than the one preceding it." If, he suggested, "the point of the entire computer revolution was to try to get a frog across a road and

b See <https://dadgum.com/halcyon/BOOK/HARRIS.HTM>

stream without being either run over by trucks or eaten by crocodiles, then it's not only unsurprising that the hacker ethic died; it isn't even sad."

That is not entirely fair, but *Frogger* does make an odd end to the main story. Levy refused to judge hackers for failing to shower, but he did not hesitate to condemn them for selling out their values. He was once a writer for *Rolling Stone*, and that attitude mirrors the culture of old-school music journalism in which beloved artists were expected to disdain commercialism while selling millions of albums. Rock journalists sneered at record labels and their besuited executives for their unseemly interest in making money and vilified them for placing constraints on artistic freedom.

The "Last True Hacker"

Levy finishes with an epilogue on Richard Stallman, who had recently launched an apparently quixotic effort to implement a free version of the hacker-friendly UNIX operating system. Stallman worked at MIT's AI lab in the 1970s but became lonely when fellow hackers left to build and sell specialized LISP workstations.

The chapter's title, "The Last of the True Hackers," gives an idea of how likely Levy thought Stallman was to succeed in his effort (or even to recruit a successor). Yet within a decade, software produced by Stallman's GNU project and Linus Torvald's work to replicate the Unix kernel had begun to challenge commercial versions of Unix. The GNU project pioneered a new model of software licensing that protected the rights of users to adapt and redistribute the software for their own needs, replicating key aspects of the original hacker culture. By the early 2000s, free software was eclipsing commercial rivals in crucial areas such as Web browsers and servers, database management systems, and programming platforms. Dominant operating systems such as Google's Android platform are built on top of free software. A broader open culture movement, similarly inspired by the hacker ethic, has produced essential resources such as Wikipedia.

Levy's decision to end the book with Stallman, drawing a direct line from the MIT hacker community to today's

world of free and open source software, has held up much better than his premature suggestion that commercialism had killed the hacker dream. In the 25th anniversary edition of *Hackers* he acknowledged that “Stallman’s fear that he would become like *Ishi, the last Yahi* was not realized.” Instead, he observed, some of the ideas in the hacker ethic “now seem so obvious that new readers may wonder why I even bothered writing them down.”¹⁰

Levy did not just capture hacker culture; he spread it to many who would never set foot in MIT, Stanford, or the Homebrew Computer Club. Since Levy wrote his book, hacker culture has become far more visible thanks to the success of the free software movement and related open culture projects such as Wikipedia. These inspired anthropological and sociological studies by scholars such as Gabriella Coleman and Christopher Keltz. Others have made broad claims for hacking as an activity central to the modern world. McKenzie Wark, for example, issued *A Hacker Manifesto* which posited the emergence of a hacker class and mimicked the 1848 *Communist Manifesto* in its call for hackers to rise up against the oppressive “vectoralists” of capitalism.¹³

The mainstreaming of hacker culture may have changed the character of computer science itself. The proportion of computer science students who were female rose steadily from the field’s beginnings in the 1960s until 1985, when it began a precipitous fall even as women’s participation in other science and engineering disciplines continued to rise.^c That is a few years after typical first experience of computing shifted from a tool in an academic context to a recreational home device for videogame playing and hacking. Computer scientists began to complain that the minds of incoming students were now contaminated by exposure to undisciplined programming methods. As Edsger Dijkstra memorably put it, “It is practically impossible to teach good programming

The original hackers were neither destructive nor dedicated to the pilfering of proprietary data, unlike the online vandals and criminals who later appropriated the word.

to students that have had a prior exposure to BASIC: as potential programmers they are mentally mutilated beyond hope of regeneration.”⁵ Correlation is not causation, but it certainly seems plausible that images of bedroom hackers and victims of “micromania” created a polarizing association of computing with a new and distinctive form of masculinity. And, as Levy himself acknowledged, classic hacker culture neither attracted nor accommodated young women.

Hacker Hybrids

Although the free and open source software movements are thriving, Levy, correctly I think, notes that the influence of hacker culture now undergirds a “world where commerce and hacker were never seen as opposing values.” In the 2010 epilogue he chatted amiably with hacker nemesis Bill Gates and mentioned that a Google executive had credited the book with inspiring his entire career. Back in the 1980s, Turkle had complained that MIT’s hackers refused to grow up and join the “real world.” As Levy showed in his 2011 book *In the Plex*, a closely observed study of Google, tech giants have created amenity filled Never Lands where hacker-infused cultures reign unchallenged and nobody ever has to grow up.¹² Levy’s titular insistence that hackers were the true “heroes of the com-

puter revolution” was echoed thirty years later when Walter Isaacson titled his blockbuster history *The Innovators: How A Group of Hackers, Geniuses, and Geeks Created the Digital Revolution*. Where Levy had described a marginal, usually overlooked subculture of computing, Isaacson’s title implied that the innovators who created personal computers, the Internet, and the modern tech industry, many of them spectacularly wealthy, famous, and powerful, somehow composed a single geeky group of hackers.

In preparing a new overview history of computing, I found myself quoting and referencing Levy more often than any other writer—not just *Hackers* but also his writing about the Macintosh project, the iPod, VisiCalc, Google, and Facebook. Levy’s great strength is his focus on people. Time after time, he has delivered the most closely observed accounts of the most important tech companies. Levy is to computer companies what veteran political journalist Bob Woodward is to presidential administrations. And like Woodward, he receives that insider access because his sources have a reasonable expectation that they will be portrayed sympathetically.

Levy’s writing suggests he is drawn to smart, eccentric, and creative people. He wants to see the best in them. In *Hackers* that works wonderfully, because his sympathetic gaze was turned on obscure characters who might be harshly judged by casual observers. He described the blinkered world view of his characters, their lack of respect for regulations and institutions, and their conviction that nothing should stand between them and the possibilities opened by new technologies. When the hackers saw something that struck them as inefficient or illogical they would go ahead and redesign it, without seeking permission or investigating alternative perspectives. That is not so far from Facebook’s motto of “move fast and break things,” or from an admonition of Zuckerberg’s that Levy quotes in the introduction to his recent *Facebook: The Inside Story*: “think of every problem as a system. And every system can be better.”⁹

The blending of hacker culture with big tech dominance concerns me, because a mind-set that might

c For a recent summary of the huge literature on this topic, see Misa, Thomas J. “Gender Bias in Computing.” In *Historical Studies in Computing, Information, and Society*. William Aspray, Ed., Cham, Switzerland: Springer Nature, 2019, 113–133.

INTERACTIONS



ACM's *Interactions* magazine explores critical relationships between people and technology, showcasing emerging innovations and industry leaders from around the world across important applications of design thinking and the broadening field of interaction design.

Our readers represent a growing community of practice that is of increasing and vital global importance.



To learn more about us, visit our award-winning website <http://interactions.acm.org>

Follow us on Facebook and Twitter



To subscribe: <http://www.acm.org/subscribe>

Association for Computing Machinery



seem harmless, or at worst narrowly self-destructive, in obsessive systems programmers is more worrying in executives running the world's most powerful corporations. Consider the unchecked personal power of Peter Thiel, Elon Musk, or Mark Zuckerberg to circumvent government regulation, manipulate the legal system, or short-circuit the democratic process. Indeed, when writing about Facebook Levy was forced by changes in public opinion and the stirring of his own conscience to treat the firm and its leaders far more harshly than he had Google a decade earlier. The “determined solipsism” that engulfed the 9th floor of MIT's Tech Square back in 1968 now hangs like a dense cloud over much of Silicon Valley. Perhaps we need some different heroes.

Further Reading

► Pretty much anything by Levy is worth reading. My personal favorite is his 1984 essay “A Spreadsheet Way of Knowledge” (republished as <https://www.wired.com/2014/10/a-spreadsheet-way-of-knowledge/>), a deeply perceptive appreciation of the early impact of spreadsheet software. If you enjoyed *Hackers* because it shows eccentric men doing stubbornly creative things in academic environments then you are more likely to enjoy his follow-ups *Artificial Life: The Quest for a New Creation* (Pantheon, 1992) and *Crypto: How the Code Rebels Beat the Government, Saving Privacy in the Digital Age* (Viking, 2001) than his later books on Apple, Facebook, and Google.

► I already mentioned the work of Chris Kelty and Gabriella Coleman on hacker culture. But if you are interested in how the other sense of hacker, the online vandal or data thief, came to predominate, then there are several key books from the 1980s that helped to spread the new meaning to a world still unfamiliar with online communication. These include *Out of the Inner Circle* by Bill Landreth and Howard Rheingold (Microsoft Press, 1985) and *The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage* by Clifford Stoll (Doubleday, 1989).

► Law professor and activist Lawrence Lessig played a crucial role in broadening the free software movement to a more general free culture

movement. His classic contribution, *Code: and Other Laws of Cyberspace* (Basic Books, 1999) remains readable and provocative.

► Levy stressed the playful nature of early hacker culture. A similar sensibility drove cult 1979 favorite *Gödel, Escher, Bach: An Eternal Golden Braid* (Basic Books, 1979), by physicist turned cognitive scientist Douglas Hofstadter. Although Hofstadter denies any personal interest in computers, his book showcases the hacker fondness for word play, recursion, baroque music, mathematical codes, and the manipulation of symbols. □

References

1. Anonymous. Business Week reports to readers on: Computers. *Business Week* 21 (June 1958).
2. Brand, S. Spacewar: Fanatic life and symbolic death among the computer bums. *Rolling Stone* (Dec. 7, 1972), 50–58.
3. Cringely, R.X. *Accidental Empires: How the Boys of Silicon Valley Make their Millions, Battle Foreign Competition, and Still Can't Get a Date*. Addison-Wesley, Reading, MA, 1992.
4. Data Processing Management Association, Executive Committee Meeting Minutes, Aug. 5–6, 1966, contained in *Data Processing Management Association Records* (CBI 88), Charles Babbage Institute, University of Minnesota, Minneapolis.
5. Dijkstra, E.W. EWD 498: How do we tell truths that might hurt. In *Selected Writings on Computer Science: A Personal Perspective*. Edsger W. Dijkstra, Ed. Springer-Verlag, New York, 1982.
6. Freiberg, P. and Swaine, M. *Fire in the Valley: The Making of the Personal Computer*. Osborne/McGraw-Hill, Berkeley, CA, 1984.
7. Isaacson, W. *Steve Jobs*. Simon & Schuster, New York, NY, 2011.
8. Lehmann-Haupt, C. Hackers as heroes. *New York Times* (Dec. 24, 1984); <https://nyti.ms/3uapqyn>
9. Levy, S. *Facebook: The Inside Story*. Blue Rider Press, New York, 2020.
10. Levy, S. *Hackers*. O'Reilly, Sebastopol, CA, 2010.
11. Levy, S. *Hackers: Heroes of the Computer Revolution*. Anchor Press/Doubleday, Garden City, NY, 1984.
12. Levy, S. *In the Plex: How Google Thinks, Works, and Shapes Our Lives*. Simon & Schuster, New York, NY, 2011.
13. McKenzie W. *A Hacker Manifesto*. Harvard University Press, Cambridge, MA, 2004.
14. Moritz, M. *The Little Kingdom: The Private Story of Apple Computer*. William Morrow, New York, NY, 1984.
15. Platt, C. and Langford, D. *Micromania: The Whole Truth about Personal Computers*. Sphere, London, 1984.
16. Turkle, S. *The Second Self: Computers and the Human Spirit*. Simon and Schuster, New York, NY, 1984.
17. Turner, F. *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. University of Chicago Press, Chicago, IL, 2006.
18. Weizenbaum, J. *Computer Power and Human Reason: From Judgment To Calculation*. W.H. Freeman, San Francisco, CA, 1976.

Thomas Haigh (thomas.haigh@gmail.com) is a professor of history at the University of Wisconsin—Milwaukee and a Comenius Visiting Professor at Siegen University. He is the author, with Paul Ceruzzi, of *A New History of Modern Computing* to be published by MIT Press later this year. Learn more at www.tomandmaria.com/tom.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 262513311 - SFB 1187 Media of Cooperation.

Copyright held by author.

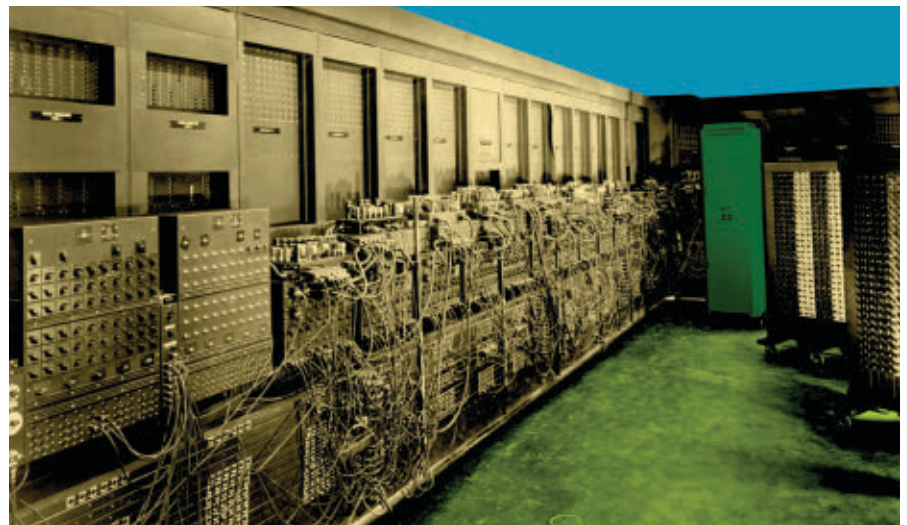
Viewpoint

Roots of ‘Program’ Revisited

Considering the fundamental nature and malleability of programming.

TODAY, IT IS a widely accepted thesis amongst historians and computer scientists that the modern notion of computer programs has its roots in the work of John von Neumann. This is symptomatic of a general tendency amongst academic computer scientists to search for the foundations of their field in logic and mathematics and, accordingly, also its historical roots. This is a distorted view of what happened: at best, the modern computer was driven by concerns of applied mathematics and developed by a collective of people (mathematicians, engineers, physicists, (human) computers, and so forth). We will not repeat why, in computing, history is reshaped in function of disciplinary identity.^{2,15} Instead, we will revisit the origins of the word “program” and argue for the need of a deeper historical understanding, not just for the sake of academic history, but for the sake of the field itself.

The notion of “program” is a fundamental one. In the flux of historical time and space, “program” underwent significant changes and has different connotations today when compared to the 1950s. Indeed, today, other words are often used instead: “software,” “apps,” or “algorithms” (as in “ethics of algorithms”). Moreover, “program” means different things to different people: a logically minded computer scientist will have a different understanding than a software engineer.



Nonetheless, as soon as one starts to speak about the historical origins of the term, this plurality of meanings disappears to be replaced by only one, namely, the “stored program.” This is anchored in another historical narrative: the modern computer originates in the “stored-program” computer. While this latter notion has been historically scrutinized,⁷ the origins of “program” have hardly been looked at independently of that notion.^a So what is the classical story here?

A Narrative

In the mid-1940s, a group of engineers of the Moore School of Electrical

Engineering, led by John Mauchly and Presper J. Eckert, designed and constructed ENIAC, a large-scale and high-speed machine that would become one of the first computers. Originally, it was a parallel and electronic machine with loops and conditionals, and could, essentially, compute any problem provided that its memory would have been unlimited.^b However, unlike some other large-scale calculators of the time, like the relay-based ASCC/Harvard Mark I or the Bell Lab machines, problems were not set up via coded instructions on punched cards or tape but were directly wired

a It should be noted that recently the notion of programmability was investigated for the Colossus.⁹

b For instance, in Eckert et al.:⁴ “[t]here is no essential or fundamental restriction imposed by the ENIAC design on the character or complication of the problems which it can do.”

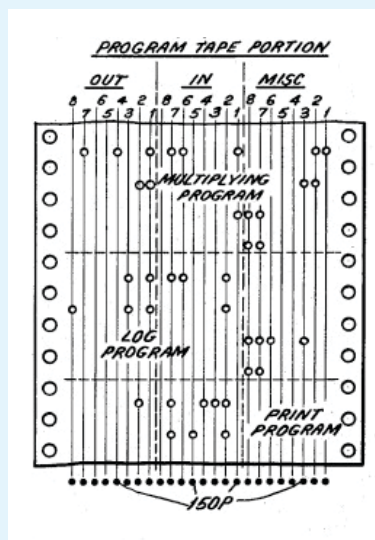
on the machine. In a sense, one had to reconfigure the whole machine every time it had to compute another problem. By consequence, setting up a problem was a time-consuming and error-prone process. In order to deal with such efficiency issues, ENIAC was converted to emulate a “stored-program” machine. However, unlike ED-SAC for instance, instructions could not be modified since programs were executed from preset switches (or, alternatively, punched cards).

It is here that von Neumann enters the story. A few months after he got involved with ENIAC, in the spring of 1945, he wrote the famous “The first draft of a report on the EDVAC,” which is considered the blueprint of the modern computer. It is then often assumed that the first “modern” programs must be those that ran on EDVAC-like machines, that is, machines like the converted ENIAC.^{5,7,8} This goes hand-in-hand with the idea that the roots of our modern conception of program should be sought with von Neumann.

But of course, just as “code” was already in use before it was introduced in the computing context (for example, Morse code), “program” (or in British spelling “programme”)^c also was an already existing word. No new term was invented at the time. It was used as a generic term to refer to a planned series of future events: an advance notice, an itinerary, something that is written before some activity happens, orders it and, so, pre-(in)scribes (programma) it. Typical examples are: a theater program, a training program, a research program, or a production plan.^{6,7} This notion was then picked up in the context of radio broadcasting to refer to radio programs. Presupposedly, Mauchly transposed this common term to the more specific engineering discourse of ENIAC and it was only with the introduction of “stored-programs” that the term really gained its current meaning.^{5,7,8} Von Neumann himself, however, hardly really used

c Note that in contemporary American spelling, “program” refers both to computer programs and, say, radio programs. In the British spelling, an explicit distinction was made between “programs” to refer to computer programs, and “programme” as used in the context of radio or theater programs.

Graphical representation of the “program tape” from the ASCC/Mark I patent, Lake et al. 1945, U.S. patent number 2,616,626.



it—he preferred the terminology of preparing, planning, setting up, and coding a problem.

Deconstructing the Narrative

We have found that “program” was already part of an extensive engineering discourse, going beyond that mentioned in the existing literature. First of all, in the context of radio engineering, the growing complexity of the broadcasting network increased the need for automation. This concerned in particular the scheduling of radio programs on different networks for different stations at different times and which had to be handled at so-called “switching points.” This resulted in a discourse in which “program” steadily transposed from radio programs to the technology itself and so one sees the emergence of terms like “program circuits,” “program trunks,” “program switching,” “program line,” “program loop,” “program selection,” and so forth.^d These are exactly the kind of terms appearing in the ENIAC context.

More importantly, we have found there is yet another engineering discourse originating in so-called “programme clocks” or “program clocks,” a device first developed in the 19th century and used to “furnish a convenient and practical clock, that may be set to strike

d See for instance U.S. patents number 2,198,326 and 2,238,070.

according to any required programme.”^e This was very handy, for instance, for a factory work floor, railway stations, or a school. In other words, they automated time schedules and production plans. From the first clocks onward, one sees the steady development of a more general technology of “program devices” or “program machines” used in a variety of applications: a paper-cutting machine, a washing machine, a calculator, and so forth.^f Here “program” comes to stand for the automatic carrying out of a sequence of operations or as an automated scheduler.

This technology came to be used also for calculating machines in the late 1930s and early 1940s, for instance, in the context of the IBM ASCC/Harvard Mark I machine. This electromechanical large-scale calculator is mostly associated with Howard Aiken, a Harvard physicist, but was designed and built by IBM engineers. The operations of that machine were controlled by the “control tape” where the sequences of operations were coded with punched holes. But while “control tape” was the standard term used once the machine was put into operation at Harvard, the original IBM patents show traces of another terminology where the control tape was also called a “program tape” and where the sequences of operations, at some points, were called “programs” instead of “sequences” (see the figure here). This terminology is due to the IBM engineers involved with the design of the machine, notably, James W. Bryce and Claire D. Lake. In fact, Bryce already had a patent in which a “program device” was introduced (U.S. patent number 2,244,241) that was capable of automatic transfer of control and other operations. Also for the ENIAC, Mauchly’s original short proposal for an “electronic computer” refers to a “program device.”¹² It is from there that the term in ENIAC developed.

Some have claimed that earlier uses of “program” in relation to ENIAC were much more restricted and referred only to specific programming circuitry in a (control) unit.⁸ This does not take into account this more general discourse which, by that time, had become

e See U.S. patent number 98678.

f See for instance U.S. patents numbers 2,134,280 or 2,026,850.

common among engineers working on automatic control, both within and outside the context of large-scale calculators. This explains why, in ENIAC, “program” had different semantic extensions and referred to individual (control) units (as in “program switches”); smaller pieces of an entire program (as in “program sequences”); or the complete schedule that organizes program sequences (as in a “complete program [for which] it is necessary to put [the] elements together and to assign equipment in detail”).¹ “Program” then refers to how automatic control, locally or globally, is organized. The semantics of the “program device” discourse is still at play here, but generalizes from the sequencing of operations to include also the scheduling of sequences of operations.

This is still evident in Hartree’s later definition (1949) of “programming” where this notion is used with reference to any “large automatic digital machine”: “programming is the process of drawing up a schedule of the sequence of individual operations required to carry out the calculation.”¹⁰ The main difference between programming ENIAC before and after its conversion to an EDVAC-like machine is that in the second case the set-up is automated through “a 100-way switch”¹⁰ where each position of the switch corresponded to a different “computing sequence.”⁸

To put it differently, the general understanding of “program” was first grafted onto the existing discourse on program devices, not on specific techniques for implementing them. Whether a program uses coded instructions stored externally on cards or tape, or, internally, on flipflops or other circuits, or whether a program is set up by wiring plugboards or by flipping switches is, from that perspective, non-essential for the meaning of “program.”

g Of course, subsequent programming practices would impact later definitions. So, for instance, in the EDSAC, an EDVAC-like and so serial machine with a symbolic assembly system, the definition of programs shortened to: “A sequence of orders for performing some particular calculation.”¹⁶ In other contexts, where flowcharting played an important role, emphasis was more on the planning aspects of programming, partially referencing back to earlier practices of human computation (see the 1954 ACM Glossary by Grace Hopper).

The “stored program” that will become commonplace later, is but one specific technique for materializing a program, that of storing a coded program internally in the computer. This, in a sense, should not be surprising: While “programs” were very much determined and dependent on the computational technology on which they are ultimately implemented and ran, that need not mean that understandings of “program” should be reduced to a specific technology. If we would have done that, we would have never had, say, concurrent programs, virtual machines or Docker containers.

Why This Matters

The 2012 Turing centenary made clear the academic computing field tends to construct a storyline where the presumed theoretical foundations of the field coincide with its historical foundations (the “first” computers and the “first” programs). This strengthens a computing discipline where one often cares more about formalism than about actual programming¹³ and contributes to a growing “communication gap” between different communities. This affects not just research and education policies, but also how we understand this field we call computing.³

As we showed, “program” did not coincide with the “stored-program” concept, rather it naturally evolved from an engineering context. Program devices for automatic control of operations were developed first for scheduling activities or communications, but were then applied to computing machines as well. In this context, a transfer of meaning happened, preparing the ground for our modern notions of program. Should we derive from this that, actually, computing should be understood first of all as an engineering discipline? No. This would miss out on the subsequent evolution of the term, when it met up again with practices of coding, of planning (manual or mechanical) calculations or of industrial process planning. If one confines oneself to one perspective only, one lacks a basic understanding: computing is not mathematics, it is not engineering, it is not logic, it is not science, it is not process control but a field on its own and one which should, perhaps, not be reduced to the confines of

disciplinary thinking (which is, itself, a construction of the 19th century). Abiding by such confinement may lead to errors and failed opportunities, as Hennessy and Patterson pointed out¹¹ with respect to software design and hardware architecture.

“[W]hen experience is not retained [...] infancy is perpetual. Those who cannot remember the past are condemned to repeat it”¹⁴ History can and has been used to reinforce confines but it can also be used against them. We must not see our historical legacy as a burden, but as the natural environment to think about the future. ■

References

- Curry, H.B. and Wyatt, W. A study of inverse interpolation on the Eniac, Aberdeen Proving Ground, Maryland, Report nr. 615, 19 August 1946.
- Bullyncck, M., Daylight, E.G., and De Mol, L. Why did computer science make a hero out of Turing? *Commun. ACM* 58, 3 (Mar. 2015), 37–39.
- Denning, P.J. and Tedre, M. *Computational Thinking*. MIT Press, 2019.
- Eckert, P.J. et al. Description of the ENIAC and comments on electronic digital computing machines. Contract W 670 ORD 4926, Moore School of Electrical Engineering, University of Pennsylvania, November 30, 1945.
- Grier, D.A. The ENIAC, the Verb ‘to program’ and the emergence of digital computers. *IEEE Annals for the History of Computing* 18, 1 (Jan. 1996), 51–55.
- Grier, D.A. Programming and planning. *IEEE Annals for the History of Computing* 33, 1 (Jan. 2011), 85–87.
- Haigh, T., Priestley, M., and Rope, C. *ENIAC in Action. Making and Remaking the Modern Computer*. MIT Press, 2016.
- Haigh, T. and Priestley, M. Where code comes from: Automatic Control from Babbage to Algol. *Commun. ACM* 59, 1 (Jan. 2016), 39–44.
- Haigh, T. and Priestley, M. Colossus and programmability. *IEEE Annals for the History of Computing* 40, 3 (Mar. 2018), 5–27.
- Hartree, D.R. *Calculating Instruments and Machines*. University of Illinois Press, Urbana, IL, 1949.
- Hennessy, J.L. and Patterson, D.A. A new golden age for computer architecture. *Commun. ACM* 62, 2 (Feb. 2019), 48–60.
- Mauchly, J. The use of high speed vacuum tube devices for calculating. Moore School of Electrical Engineering, University of Pennsylvania, August 1942.
- Noble, J. and Biddle, R. Notes on postmodern programming. *ACM SIGPLAN Notices* 39, 12 (2004), 40–56.
- Santayana, G. *Reason in Common Sense, Volume I of The Life of Reason*. Charles Scribner’s Sons, New York, 1905.
- Tedre, M. *The Science of Computing: Shaping a Discipline*. CRC Press, 2014.
- Wilkes, M.V. and Wheeler, D.J., Gill, S. *The Preparation of Programs for an Electronic Digital Computer*, second edition. Addison-Wesley, 1957.

Liesbeth De Mol (liesbeth.de-mol@univ-lille.fr) is a researcher with the Centre National de la Recherche Scientifique and is affiliated with UMR 8163 Savoirs, Textes, Langage at the University of Lille. She is also the PI of the PROGRAMme project.

Maarten Bullyncck (maarten.bullyncck@univ-paris8.fr) is an associate professor at the departement of mathematics and history of science of Université Paris 8 and associate researcher at laboratory UMR 8533 IDHE.S.

This research was supported by the ANR PROGRAMme project ANR-17-CE38-0003-01.

Copyright held by authors.

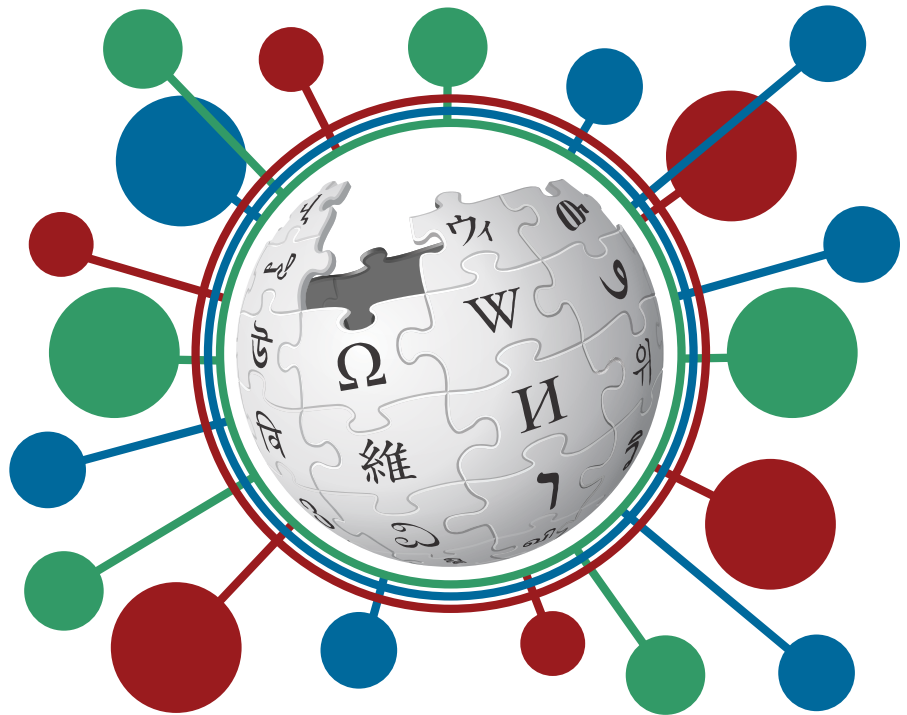
Viewpoint

Building a Multilingual Wikipedia

Seeking to develop a multilingual Wikipedia where content can be shared among language editions.

WIKIPEDIA HAS MORE than 50 million articles in approximately 300 languages. The content in these languages is independently created and maintained. The knowledge in Wikipedia is very unevenly distributed over the languages: some languages have more than a million articles, but more than 50 languages have only a few hundred articles or less. More importantly, also the number of contributors is very unevenly distributed: English Wikipedia has more than 418,000 contributors, the second-most active one, Spanish, drops down to 90,000. More than half of language editions have fewer than 10 contributors doing more than four edits per month. To assume that fewer than 10 active contributors can write and maintain a comprehensive encyclopedia in their spare time is optimistic at best.

In order to close these knowledge gaps we are building a multilingual Wikipedia where content is created only once but made available in all languages. The multilingual Wikipedia has two main components: **Abstract Wikipedia** where the content is created and maintained in a language-independent notation, and **Wikifunctions**, a project to create, catalog, and maintain functions. For the multilingual Wikipedia, the most important function is one that takes content from Abstract Wikipedia and renders it in natural language, which in turn gets integrated into Wikipedia proper.



This will considerably reduce the effort required to create a comprehensive and maintain a current encyclopedia in many languages. It will allow more people to share more knowledge in more languages than ever before. It will be particularly useful for underserved languages, providing an important way to help improve education and ready access to knowledge in many countries.^a

^a This Viewpoint provides a brief summary of the full architecture available online.⁹

Example

We follow a toy example. It does not cover the complexity of the problem space, but is used to sketch the architecture. Taking the following two (simplified) sentences from English Wikipedia:

“San Francisco is the cultural, commercial, and financial center of Northern California. It is the fourth-most populous city in California, after Los Angeles, San Diego and San Jose.”

Figure 1 shows how the text could be represented as abstract content. The example shows a single **constructor** of

type `Article` with a single **key** called `content`. The **value** of `content` is a list with two constructors, one of type `Instantiation`, the other of type `Ranking`. The `Instantiation` constructor has two keys, `instance` and `class`, where the `instance` key has a simple entity as the value, *San Francisco*, and the `class` key has a complex value built from another constructor. *San Francisco* refers to an **item** from the Wikidata catalog of items, Q62.^b Wikidata is a sister project of Wikipedia, an open knowledge base that anyone can edit,¹⁰ which currently provides language-independent identifiers for 90 million entities, such as Q62 for San Francisco, and one billion machine-readable facts about these entities. Wikifunctions will be able to call Wikidata to request these facts and use them to enrich content.

We require one **renderer** per constructor and language. A renderer is a **function** that takes abstract content and a language and turns it into natural language text (or an intermediate object for another renderer). Renderers are created and maintained by the community.

Desiderata

Content has to be editable in any language. Note this does not mean we need a parser that can read arbitrary input in any language. A form-based editor could be sufficient and easy to localize.

The set of constructors has to be extensible by the community. We cannot assume we can create all necessary constructors to capture Wikipedia a priori.

Renderers have to be written by the community. This does not mean that every community member must be able to write renderers. Wikipedia and Wikidata have shown that contributors with different skill sets can successfully work together to tackle very difficult problems.¹

Lexical knowledge must be easy to contribute. Rendering will require large amounts of lexical knowledge. Wikidata has been recently extended to express and maintain lexical knowledge.^c

Content must be easy to contribute.

Content will constitute the largest part of the system. Accordingly, the user experience for creating and maintaining content will be crucial to the success of the project (see Figure 2).

Graceful degradation. The different languages will grow independently from each other at different speeds. It

is important the system does not stop rendering the whole article because of a single missing lexicalization.

Architecture

The community of a language Wikipedia can choose to use the abstract content that is stored alongside the items in

Figure 1. An example abstract content of two sentences describing San Francisco. The names of the constructors and their keys are given in English here, but that is merely convenience. Just as the items in Wikidata, they will be represented by language-independent identifiers.

```
Article(
  content: [
    Instantiation(
      instance: San Francisco (Q62),
      class: Object_with_modifier_and_of(
        object: center,
        modifier: And_modifier(
          conjuncts: [cultural, commercial, financial]
        ),
        of: Northern California (Q1066807)
      )
    ),
    Ranking(
      subject: San Francisco (Q62),
      rank: 4,
      object: city (Q515),
      by: population (Q1613416),
      local_constraint: California (Q99),
      after: [Los Angeles (Q65),
             San Diego (Q16552),
             San Jose (Q16553)]
    )
  ]
)
```

Figure 2. Mock up of the user interface.

On the left is a free text input box. That text is classified in order to offer a first pass of the abstract content in the top right corner, where the constructor and the respective values are being displayed and made available for direct editing. The bottom right shows a rendering of the currently edited content in a selection of languages the contributor understands. This provides feedback that allows the contributor to adapt to the system's constructor library.

<p>San Francisco is the cultural, commercial, and financial center of Northern California. <i>It is the fourth-largest city by population in California, after Los Angeles, San Diego and San Jo</i>^{se}.</p>	<p>Constructor subject San Francisco Q62 object city Q515 local constraint California Q99 temporal constraint [add] rank 4 by population size Q1613416 after Los Angeles Q65 San Diego Q16552 San Jose Q16553 [add value]</p> <p>[add other slot]</p> <p><i>English</i> [change language] San Francisco is the cultural, commercial, and financial center of Northern California. It is the fourth-most populous city in California, after Los Angeles, San Diego, and San Jose.</p> <p><i>German</i> [change language] San Francisco ist das kulturelle, kommerzielle und finanzielle Zentrum Nordkaliforniens. Es ist, nach Los Angeles, San Diego und San Jose, die viergrößte Stadt in Kalifornien.</p> <p>[add language]</p>
--	---

^b <https://www.wikidata.org/entity/Q62>

^c https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

Distinguished Speakers Program

A great speaker can make the difference between a good event and a WOW event!

Students and faculty can take advantage of ACM's Distinguished Speakers Program to invite renowned thought leaders in academia, industry and government to deliver compelling and insightful talks on the most important topics in computing and IT today. ACM covers the cost of transportation for the speaker to travel to your event.

speakers.acm.org



Association for
Computing Machinery

Wikidata. Natural language text is generated from abstract content using the render function available in Wikifunctions and then displayed in Wikipedia to cover current gaps. Content from Abstract Wikipedia and locally created natural language content will be living side by side. The abstract content in Wikidata and the renderers in Wikifunctions are composed from constructors and functions created and maintained in Wikifunctions. The functions can call the lexicographic knowledge in Wikidata, for example, for irregular inflections. This architecture is sketched in Figure 3. The constructor specification states the type of the result of the specification when being rendered. This allows for a system built on the principles of functional programming, which has proven suitable for natural language generation.⁶

For every function, type, and constructor, there is a page in Wikifunctions with their definition and documentation, their keys, whether the keys are optional, and what type of values are allowed for each key. The most relevant function in Wikifunctions with regards to Abstract Wikipedia is a function to render abstract content in natural language. Wikipedia calls this rendering function with some content and the language of the given Wikipedia, and displays the result as article text.

I glossed over many issues such as agreement, saliency, or register. Wikifunctions will need to implement a natural language generation library⁷ taking these issues into account.

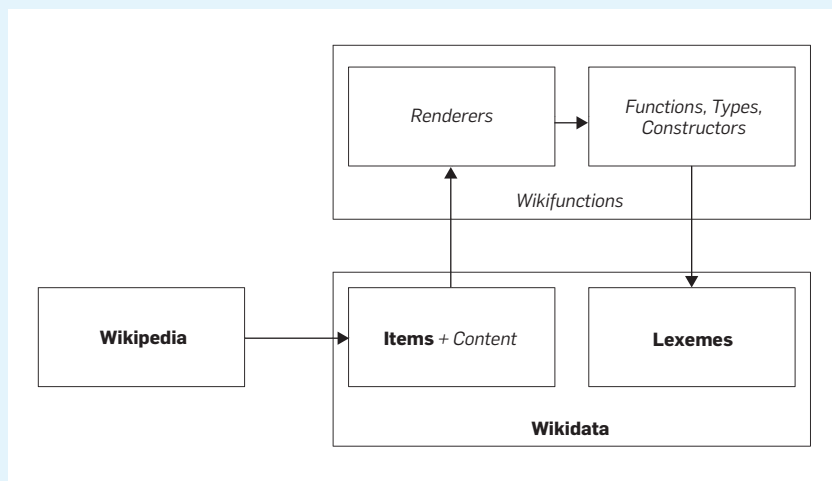
Wikifunctions

The primary goal of Wikifunctions is to enable the multilingual Wikipedia: define the constructors for abstract content and implement the rendering function. This will lead to the creation of an open, widely usable, and well-tested multilingual natural language generation library.

But Wikifunctions also has a secondary goal: to provide a comprehensive library of functions, to allow everyone to create, maintain, run, and reference functions. This will enable people without a programming background to compute answers to many questions, either through Wikifunctions or through third-party sites accessing the functions. It offers a place where scientists collaboratively create models. It provides a persistent identifier scheme for functions, thus allowing processes, scientific publications, and standards to refer to functions unambiguously. New programming languages will be easier to create as they can rely on Wikifunctions for a comprehensive library. One obstacle in the democratization of programming is

Figure 3. Architecture of the multilingual Wikipedia proposal.

To the left is the Wikipedia language edition. The Wikipedia language edition calls the content stored in Wikidata and the content is then rendered by renderers in Wikifunctions. The result of the rendering is displayed in Wikipedia. The renderers are functions in Wikifunctions, and there they can use other functions, types, and constructors from Wikifunctions. These function finally can call the lexicographic knowledge stored in Wikidata. In italics are the parts that are proposed here (Content, Wikifunctions), in bold the parts that already exist (Wikipedia, Wikidata without Content).



that most programming language are based on English.³ Wikifunctions will use language-independent identifiers, allowing to read and write code in any natural language.

Function specifications can have multiple implementations. Implementations can be in a programming language such as JavaScript, WebAssembly, or C++, or composed from other functions. Evaluators can execute different implementations and compare their results and runtime behavior. Evaluators can be implemented in a multitude of backends: the browser, the cloud, the servers of the Wikimedia Foundation, on a distributed peer-to-peer evaluation platform, in a mobile app, or on the user's machine.

Function calls to Wikifunctions can be embedded in several contexts. Wikifunctions will provide UIs to call individual functions, but it also lends itself to be used from a Web-based REPL, locally installed CLIs, a RESTful API, as a library imported in a programming language, through dedicated apps, Jupyter notebooks, natural language assistants, or spreadsheet cells.

Risks and Advantages

Leibniz was probably the best-known proponent of a universal language, the *characteristica universalis*. Such ambitions have repeatedly failed.⁴ The main difference to Abstract Wikipedia is that Leibniz not only aimed for a notation for knowledge but also for a calculus to derive veracity; here, the focus is solely on notation.

A major risk is that contributing to Abstract Wikipedia and Wikifunctions becomes too difficult. Like all Wikimedia projects it relies on a sufficient number of contributors. But Wikimedia communities have repeatedly tackled very hard tasks. They managed to self-organize and allow people with different skillsets to collaborate, and to succeed beyond expectations on projects such as an encyclopedia¹ or a knowledge base.¹⁰ It will be crucial to provide an accessible user experience.

A major risk is that the number of constructors is too high. If the number of constructors remains in the low thousands, a community of approximately five contributors can unlock a current and comprehensive encyclope-

Wikifunctions and Abstract Wikipedia are expected to drive a number of research directions in knowledge representation, natural language generation, collaborative systems, and computer-aided software engineering.

dia for their language. Coverage experiments on texts⁵ using FrameNet² allowed us to be optimistic about this, but the results are preliminary. There are several reasons to be optimistic:

- ▶ We aim only at a single genre, encyclopedias.
- ▶ The exact surface text is not so important as long as the text retains fidelity.
- ▶ We start simple and allow iteration.
- ▶ We do not need natural language understanding, merely generation.
- ▶ The baseline is very low.

Wikifunctions and Abstract Wikipedia are expected to drive a number of research directions in knowledge representation, natural language generation, collaborative systems, and computer-aided software engineering. Having a large catalog of functions will be valuable for many tasks, as will the creation of a large multilingual natural language generation library.

Advances in machine learning, for example, for article generation,⁸ can be neatly tied into the architecture. ML systems for learning renderers from example texts, systems to improve the fluency of rendered text, or classifiers that help generate content from natural language input can all be important modules in the project. Machine learned components can be made ac-

cessible like any other function in Wikifunctions, or they can be used for offline analysis of Abstract Wikipedia or potential input text. Such combinations will guide an interesting exploration in the mechanisms and effectiveness of human-machine teams.

Conclusion

Building a multilingual Wikipedia is a clearly defined and highly attractive goal with many challenging problems. We invite the research community to join us. With Abstract Wikipedia and Wikifunctions, we sketch out an architecture to get there. A major advantage of splitting up Wikifunctions and Abstract Wikipedia is that it recognizes the risks in the project. Wikifunctions defines a valuable goal by creating a catalog of functions. Abstract Wikipedia provides value by improving on the maintainability of currently bot-created articles. Even if the full vision is not achieved, we identify valuable intermediate milestones. The project is achievable without the need for research breakthroughs. The current state of the art in natural language generation, knowledge representation, and collaborative systems can be tied together to create a system that enables many more people than today to share in the sum of all knowledge. ■

References

1. Ayers, P., Matthews, C., and Yates, B. *How Wikipedia Works (And How You Can Be a Part of It)*. No Starch Press, 2008.
2. Baker, C.F., Fillmore, C.J., and Lowe, J.B. The Berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998, 86–90.
3. Dasgupta, S. and Mako Hill, B. Learning to code in localized programming languages. In *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, L@S '17*, New York, NY, USA, 2017, 33–39.
4. Eco, U. *The Search for the Perfect Language (the Making of Europe)*. Blackwell, 1995.
5. Ferraro, F. et al. Concretely annotated corpora. In *Proceedings of the AKBC Workshop at NIPS*, 2014.
6. Ranta, A. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI, 2011.
7. Reiter, E. and Dale, R. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.
8. Vougiouklis, P. et al. Neural Wikipedia: Generating textual summaries from knowledge base triples. *J. Web Semant.*, 52-53 (2017), 1–15.
9. Vrandečić, D. Architecture for a multilingual Wikipedia. arXiv preprint arXiv:2004.04733, 2020; <https://arxiv.org/abs/2004.04733>.
10. Vrandečić, D. and Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57, 10 (Oct. 2014), 78–85; doi: <http://dx.doi.org/10.1145/2629489>.

Denny Vrandečić (denny@wikimedia.org) is Head of Special Projects at the Wikimedia Foundation in San Francisco, CA, USA.

Copyright held by author.

Arab World Regional Special Section



ILLUSTRATION BY SPOOKY POOKA AT DEBUT ART.
FOR CREDITS ON IMAGES IN COLLAGE, SEE P.3



Welcome

WELCOME TO THIS special section on the Arab world, covering Arab speaking countries from the Atlantic Ocean to the Gulf. The Arab world includes diverse countries with varying ethnicities, cultures, dialects, history, and socioeconomic backgrounds. The countries are also extremely diverse in the abundance of human and natural resources with a high variance of distribution of each across the Arab nations. With limited research support in many countries and a lack of a corresponding consuming industry, the Arab region has struggled in computing research till the early 2000s. Nevertheless, over the last decade, the region witnessed a significant investment in research support and the corresponding enabling industry. Several countries are moving toward a knowledge-based economy, while others have opened up to more industry. This is apparent by the recent inception of new research-based universities and labs in Egypt, Morocco, Qatar, Saudi Arabia, and United Arab Emirates, and the recent change of the computing industry landscape in Egypt, Jordan, and Lebanon.

To create this special section, we launched an open call for contributions to a virtual workshop. The workshop, held on Aug 29–30, 2020, featured 32 selected talks, five invited academic keynotes, and two invited industrial keynotes. Following the workshop, the organizing group met twice with the Advisory Board Committee members, Ahmed Elmagarmid (QCRI, Qatar), Mootaz Elnozahy (KAUST, Saudi Arabia), Lina Karam (LAU, Lebanon), and Taieb Znati (UAEU, UAE). The meetings concluded with selection of 18 papers to invite for this special section. Nine of these papers were selected from the workshop, while the rest were invited to ensure diversity and broad representation without sacrificing quality.

Articles in this issue are diverse in several aspects. Geographically, the authors represent 12 Arab countries, namely, Algeria, Egypt, Jordan, Lebanon, Morocco, Palestine, Qatar, Saudi Arabia, Sudan, Syria, Tunisia, and the UAE. Out of the 18 papers, 11 have authors from academia, nine have authors from research labs, six from the industry, one from government, and one from an NGO. Nine out of the 18 papers have female authors while 16 have male authors. In terms of topics, the papers cover AI, bioinformatics, computer networks, database systems, education, HCI, HPC, ML, NLP, security, speech, and transportation.

Of course, this issue is by no means comprehensive of all the exciting research in the Arab world. Yet, it is a start that would help strengthen the research collaboration and communication with various regions of the world. Articles in this issue are either describing a topic very specific to the region and/or distinguished research that came out of the region and had international impact. We sincerely hope that readers will find the articles exciting and eye opening. The Arab world has a huge potential in research and development that is yet to be exploited.

Finally, we would like to thank all the authors, workshop participants, and our advisory board for their efforts in coming up with this issue. Special thanks go to Christine Bassem (Wellesley College, USA), Tamer Elbatt (AUC, Egypt), and Cherif Salama (AUC, Egypt) for their tireless efforts in the workshop organization and selection of articles in this issue.

—*Sherif G. Aly, Mohamed Mokbel, and Moustafa Youssef*
Coordinators of the Arab World Region Special Section

Sherif G. Aly is a professor and chair of the Department of Computer Science and Engineering at The American University in Cairo, Egypt.

Mohamed Mokbel is Chief Scientist at Qatar Computing Research Institute, HBKU, Doha, Qatar.

Moustafa Youssef is a professor at the American University in Cairo and Alexandria University, Egypt.

Copyright held by authors/owners.

EDITORIAL BOARD

EDITOR-IN-CHIEF

Andrew A. Chien
 eic@cacm.acm.org

DEPUTY TO THE EDITOR-IN-CHIEF

Morgan Denlow
 cacm.deputy.to.eic@gmail.com

CO-CHAIRS, REGIONAL SPECIAL SECTIONS

Sriram Rajamani
 Haibo Chen
 P. J. Narayana

SPECIAL SECTION CO-ORGANIZERS

Sherif G. Aly
 The American University in Cairo

Mohamed Mokbel
 Qatar Computing Research Institute, HBKU
 Moustafa Youssef
 The American University in Cairo and Alexandria University, Egypt

SPECIAL ADVISOR

Christine Bassem
 Wellesley College



Watch the co-organizers discuss this section in the exclusive *Communications* video.
<https://cacm.acm.org/videos/arab-world-region>

Hot Topics

- 46 **Building a Research University in the Arab Region: The Case of KAUST**
By *E.N. (Mootaz) Elnozahy*
- 50 **Building a Preeminent Research Lab in the Arab Region: The Case of QCRI**
By *Ahmed Elmagarmid and Abdellatif Saoudi*
- 54 **Data Science for the Oil and Gas Industry in the Arab Region**
By *Motaz El Saban*
- 57 **The Strategic Pursuit of Artificial Intelligence in the United Arab Emirates**
By *Farah E. Shamout and Dana Abu Ali*
- 59 **An AI-Enabled Future for Qatar and the Region**
By *Ashraf Abounnaga, Sanjay Chawla, Ahmed Elmagarmid, Mohammed Al-Mannai, and Hassan Al-Sayed*
- 62 **Entrepreneurship Ecosystem in Lebanon**
By *Walid R. Touma and Saad El Zein*
- 64 **Autonomous Driving in the Face of Unconventional Odds**
By *Hesham M. Eraqi and Ibrahim Sobh*
- 67 **Traffic Routing in the Ever-Changing City of Doha**
By *Sofiane Abbar, Rade Stanojevic, Shadab Mustafa, and Mohamed Mokbel*
- 69 **ArabHCI: Five Years and Counting**
By *Shaimaa Lazem, Mennatallah Saleh, and Ebtisam Alabdulgader*

Big Trends



- 72 **A Panoramic Survey of Natural Language Processing in the Arab World**
By *Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Husein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak*
- 82 **The Arab World Prepares the Exascale Workforce**
By *David Keyes*
- 88 **Non-Traditional Data Sources: Providing Insights into Sustainable Development**
By *Ingmar Weber, Muhammad Imran, Ferda Ofli, Fouad Mrad, Jennifer Colville, Mehdi Fathallah, Alissar Chaker, and Wigdan Seed Ahmed*
- 96 **Cyber Security Research in the Arab Region: A Blooming Ecosystem with Global Ambitions**
By *Christina Pöpper, Michail Maniatakos, and Roberto Di Pietro*
- 102 **Unleashing Early Maturity Academic Innovations**
By *Slim Abdennadher, Sherif G. Aly, Joe Tekli, and Karima Echiabi*



- 108 **Biomedical Computing in the Arab World: Unlocking the Potential of a Growing Research Community**
By *SeifEldawlatly, Mohamed Abouelhoda, Omar S. Al-Kadi, Takashi Gojobori, Boris Jankovic, Mohamad Khalil, Ahsan H. Khandoker, and Ahmed Morsy*
- 114 **Networking Research for the Arab World: From Regional Initiatives to Potential Global Impact**
By *Basem Shihada, Tamer ElBatt, Ahmed Eltawil, Mohammad Mansour, Essaid Sabir, Slim Rekhis, and Sanaa Sharafeddine*
- 120 **Database Systems Research in the Arab World: A Tradition that Spans Decades**
By *Ashraf Abounnaga, Azza Abouzied, Karima Echiabi, and Mourad Ouzzani*
- 124 **Connecting Arabs: Bridging the Gap in Dialectal Speech Recognition**
By *Ahmed Ali, Shammur Chowdhury, Mohamed Afify, Wassim El-Hajj, Hazem Hajj, Mourad Abbas, Amir Hussein, Nada Ghneim, Mohammad Abushariah, and Assal Alqudah*



Association for Computing Machinery
Advancing Computing as a Science & Profession

Building a Research University in the Arab Region: The Case of KAUST

BY E.N. (MOOTAZ) ELNOZAHY

THE ESTABLISHMENT OF King Abdullah University of Science and Technology (KAUST) in 2009 was the fulfillment of a lifelong dream of its founder, the late King Abdullah of Saudi Arabia. His vision for the university was deeply rooted in the historical and cultural contexts of the Middle East. He in-

tended the university to be seen as a revival of the old “house of wisdom,” which was a premier institution of learning in Baghdad from the 9th century until the 13th century. Starting as a private library of the fabled Caliph Harun Al-Rasheed, it developed quickly into the 9th century equivalent of a research laboratory and a university. The house of wisdom was the birthplace of algebra

and was a milieu where many developments took place in various fields of science and humanities. It was sponsored generously by the Abbasid Caliphate and welcomed any knowledge seeker regardless of ethnic origin, religious beliefs, or sex.² Its rise and fall coincide respectively with the start and end of the golden age of the Islamic civilization.

The king’s vision was to establish a modern university that selectively picks from the best practices of modern, western universities. The model of the university was inspired by the California Institute of Technology (Caltech) in terms of its size and focus on STEM. The university’s original research vision sought to focus on four major challenges that face the kingdom (and humanity at large), namely, food,

water, energy and the environment.

Though KAUST is a modern university that shares many of the operational mechanisms of western universities such as faculty evaluation and promotion, degree requirements, and the like, the university has a unique model of operation that deviates from most modern universities in several important aspects:

► KAUST shares the old house of wisdom’s model of providing unconditional support to its faculty. Professors who join the university enjoy perpetual annual funding to pursue curiosity-based research. Additionally, research centers create focal points for goal-oriented research to advance knowledge toward solving big and practical problems. Individually or within centers, the

The first two professors to join the university were high-profile computational scientists, and the university procured a supercomputer even before it opened.



King Abdullah University of Science and Technology (KAUST).

researchers have at their disposal excellent infrastructure that provides anything from the oscilloscope to a supercomputer. The university offers financial support to all its students. A generous endowment enables this level of support and promotes the university's freedom of intellectual quest.

► Inspired by the international nature of the old house of wisdom, the international character of KAUST is based on a modern-day quest for making the university a bridge that connects cultures. KAUST brings people from all over the world to study, live, and work together, and with more than 114 nationalities that live on campus, KAUST takes diversity to its fullest scale compared to any university worldwide. This diversity has a strong implication on the admission of students. The cultural diversity of the student body is a first-class goal along with academic excellence. The proportions of this cultural mosaic are continuously monitored,

and often otherwise excellent students are not admitted if it would lead to a cultural imbalance within the student body. About 35% to 40% of the students are Saudis, and about 35% of the students are women, which is a healthy number for a STEM-only institution.

► KAUST has ambitious goals to spur economic development. The king's vision recognizes the importance of science and technology in creating a "knowledge economy" that can move the kingdom of Saudi Arabia from a commodity-based economy toward a more diversified one with high paying jobs. Entrepreneurship education is required of all its students and an economic cluster encourages start-up companies and technology transfer.

► KAUST is a graduate-only institute that focuses on research and grants. Master of Science and Doctor of Philosophy degrees only. Nevertheless, the university has a considerable number of research programs that are oriented

toward a few hundred undergraduate student interns throughout the year. The internships attract students from all over the world, and some of the interns end up applying for graduate studies at KAUST.

► KAUST does not have a tenure system. The origins of the tenure system in the 1920s were studied carefully along with some of its unintended consequences and it was decided that a model of rolling contracts would serve the university better.

Computer and Computational Sciences at KAUST

While the original four areas of research focus did not include computer science or computational sciences, these topics were a target of substantial investment from the beginning. In fact, the first two professors to join the university were high-profile computational scientists, and the university procured a supercomputer even before it opened. This was in recognition of the central

role that computing takes in propelling research across a broad range of scientific endeavors. These investments were matched from the beginning by the popularity among student applicants, who disproportionately apply to computer science compared to other fields (although biosciences and electrical engineering sometimes come close). The popularity of computing did not stop there—today over half the faculty at KAUST use computational research methods, and this number is expected to grow as artificial intelligence continues to make inroads in all fields of scientific research.

Like Caltech, KAUST is a small university and aims to have around 225 professors at maturity. This posed an interesting challenge in setting the strategy for computer science: Should the focus be on breadth to cover as many topics as possible, or should the focus be on a few areas of depth? Taking advantage of the absence of undergraduate education's constraints,

the decision was to follow the path of narrow depth over shallow breadth, as it was important to strive for critical mass and excellence. It was deemed important to establish the brand name of computer science at KAUST as a place where there are few areas of world class excellence. The areas of initial investments focused on biosciences and computer graphics, with two corresponding centers established as focal points. A center of extreme computing was later established in 2015.

In 2018, the appointment of Tony Chan, a prominent computational scientist, as university president signaled a transition in the university strategy. A “digital” focus was officially added to the original four, and a new strategy was drafted with a desire to establish new strengths in artificial intelligence, cybersecurity, robotics, smart health, and modeling. All these areas are within the realm of computer and computational sciences. Two centers in artificial intelligence and cybersecurity will soon be established.

Challenges

Establishing a research university is a challenge no matter where. The Middle East adds to this challenge in many ways. This article focuses on the challenges that are intrinsic to computer science, leaving out the broader challenges that face the university overall:

- ▶ *Lack of local information technology industry with a critical mass.* Most research-oriented universities enjoy a symbiotic relationship with industry, which provides problems to solve, jobs for the university graduates, and often financial support. The state of the industry in computer-related fields in the Arab world is poor.

- ▶ *Access to talent.* This is a universal problem in the information technology field, but it is more exacerbated in the Middle East. Even with generous support and lucrative compensation, the university simply cannot match the industry salaries either for young graduates (affecting potential student recruits) or experienced researchers (affecting potential postdoc, research scientist



KAUST campus in Thuwal, Saudi Arabia.

and even faculty hires). The problem manifests itself both with recruitment of talent and retaining it. This has proved to be the most challenging problem. The university grew nevertheless, although a desire to reach the maturity size in 10 years will have to wait further to be achieved, as there is no intention to compromise the quality in this regard.

- ▶ *Cultural challenges.* With few exceptions, the culture of research in the Middle East was largely dormant for the last 700 years. Today, the general population at large is impatiently waiting for results, and it may have unrealistic expectations out of research while not recognizing the necessity of ample time and resources to build the foundation of a modern university. Impatience often leads to frustration and questions about the value of the university and the underlying investment.

Perspective

A university that is an 11-year-old institution is by definition “a work in progress.” Yet, KAUST is a Saudi national project

that was launched with lofty expectations, and it is inevitable that the issue of evaluation is a frequent topic of discussions. By traditional measures, it is unquestionably too early to declare success; but one can confidently glean several positive trends that are pointing in the right direction. For instance, a recently popular measure of evaluating universities is the sort of job placement that graduates receive upon completing their degrees.³ By this measure, KAUST has made great strides for a university of its age. Among the computer science alumni are assistant professors at the Toyota Technological Institute in Chicago and the University of California at Irvine and about a couple of dozen universities most notably in Saudi Arabia and China; research scientists at U.S. national laboratories such as Oakridge and Berkeley laboratories; postdoctoral fellows at Oxford, IBM Research, among other academic institutions of high repute; and choice positions in large firms such as Google at Mountain View, NVIDIA, ARAMCO, Intel among



Inside KAUST's Visualization Laboratory.

others. Another measure of evaluating universities is how they fare in rankings, though many ranking systems are controversial because of the subjectivity inherent in the process. In computer science, the CS Ranking¹ is arguably one that uses only objective measures. In this ranking, KAUST's Visual Computing Center, which specializes in computer vision, computer graphics and visualization is ranked 12th worldwide in the period 2010–2020. If one includes all the areas in computer science at KAUST, the combined ranking comes to 37th worldwide for the same period.¹ These are

be the most relevant in the Middle Eastern context. For example, success in job placement in places like Oxford or IBM Research overseas may be an undisputable indicator of the high quality of research and education that take place at KAUST, but they also indicate that the university may be unwittingly playing a role in the hemorrhage of talent outside the region. Similarly, rankings are often nuanced and controversial, and at any rate, should never be a goal by themselves. The Middle Eastern citizen who is footing the bill for the educational system has more serious



A library on the KAUST campus.

respectable rankings for a young, small department. In light of KAUST's performance in these measures, one can argue the vectors are pointing in the right direction, but a more nuanced and closer look calls for additional ways of measuring success.

The aforementioned traditional measures of evaluating academic institutions have been borrowed from best practices in the developed world. They nevertheless may not

problems for which a good university ranking may not prove very relevant. In fact, the overindulgence with rankings may well lead to undesirable behavior even in industrialized economies.⁴ This points to the need to develop indigenous systems of evaluation that take the local context into account, instead of merely copying the western models. Of course, a full-fledged discussion of this topic is an article in and of itself. However, one can

Computer science has the lion's share of all the start-ups that were created by KAUST graduates.

list a few salient features that may be more relevant to the Middle Eastern context: The contribution to GDP and GDP growth, the number of companies that get started by young university graduates, the number of jobs created as a result of the research, and the development of indigenous new technologies that solve local problems directly and innovatively. None of these measures may be easy to compute objectively, but they will be more relevant to the local context than citations, h-index, or so-called high-impact publications. In this context, the indicators for computer science at KAUST are also positive. Several start-ups came out of the university graduates, deploying in Saudi Arabia amid a fledgling IT industry, with the university sometimes playing an investor role and sometimes an incubator. In fact, computer science has the lion's share of all the start-ups that were created by KAUST graduates. Additionally, the department has graduated many highly trained engineers and scientists who have stayed in the kingdom and contribute to the local economy, including international students who decided to stay. A more precise measure of the impact requires more work.

In closing, KAUST was the brainchild of a visionary king who wanted to revive the tenets of his culture, which places science and knowledge at the highest esteem. He established a modern-day house of wisdom as an international, research-oriented Middle Eastern university with an unusual model of funding and an unwavering commitment to diversity. Early indicators point to a successful operation and a promising future amid nontrivial challenges. The road ahead requires building on the early successes, but also there is a need to create stronger links to the local economy and to develop indigenous measures and metrics for success that are more relevant to the local context and that can highlight the contributions of the university in a manner that will be better appreciated. ■

References

1. Computer Science Rankings; <https://www.csranks.org>.
2. Lyons, J. *The House of Wisdom*. Bloomsbury Press, 2010.
3. Smith-Barrow, D. and Kerr, E. *U.S. News and World Report*, Jan. 2020.
4. Vardi, M.Y. Academic rankings considered harmful! *Commun. ACM* 59, 9 (Sept. 2016).

E.N. (Mootaz) Elnozahy is currently a special advisor to the president and a professor of computer science at King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Previously, he was the dean of the Computer, Electrical and Mathematical Sciences and Engineering Division.

© 2021 ACM 0001-0782/21/4

Building a Preeminent Research Lab in the Arab Region: The Case of QCRI

BY AHMED ELMAGARMID AND ABDELLATIF SAOUDI

THE QATAR COMPUTING Research Institute (QCRI) is one of three national research institutes established in 2010 by Qatar Foundation (QF) for education, science and community development. It operates under the umbrella of Hamad Bin Khalifa University and is steered operationally by the Research, Development, and Innovation (RDI) division, which was established within QF to oversee the three national research institutes' day-to-day operations. In this capacity, RDI provides high-level planning, coordination, and oversight to further the institutes' research priorities.

QCRI was created with

a mandate to support Qatar's transformation from a carbon economy to a knowledge-based economy. In doing so, it fulfills Qatar Foundation's overarching objectives of enabling national and regional change. QCRI's mission is to conduct innovative, multidisciplinary applied computing research that addresses national priorities that enhance citizens' quality of life, enables broader scientific discoveries, and makes local businesses more competitive globally. The Institute is focused on tackling large-scale computing challenges for growth and development relevant to Qatar, the wider Arab region, and the world. The cutting-edge research that QCRI conducts is AI-

based in Arabic language technologies, social computing, data analytics, and cybersecurity.

In the Foundation's early days, Qatar and QF leadership invited Arab expat scientists (AES) from the diaspora to be part of Qatar's burgeoning scientific renaissance. AES membership increased from a handful to almost 1,000 in a matter of years. Scientists across various fields were invited to Qatar to spend a few days as guests to learn about the country's vision. In 2006, what became known as the "Arab Expat Scientists Forum" began conducting a series of annual events that explored technical tracks within scientific disciplines. The events were intended to be broad and

inclusive, and there were no attempts to formalize the meetings beyond facilitating the exchange between local scientists and expats.

Another important undertaking was the AES's push in 2008 to create an informal organization to connect, network, and expand the reach of the group's activities in Qatar. A steering committee stemmed from the larger group, and a framework was formed to shepherd the steps required to achieve the initiative's goals. The committee spent a considerable amount of time identifying areas of urgent importance to Qatar. The process was consultative and deliberative and eventually resulted in the decision to focus on three areas:

ALL PHOTOS COURTESY OF QATAR COMPUTING RESEARCH INSTITUTE (QCRI)



Education City, where offices for Qatar Computing Research Institute are housed.

Life Sciences, Energy and Environment, and Information and Communications Technology (ICT).

Within ICT, the focus shifted to the science and engineering of computing and information. Two members of the steering committee were selected to lead the development of this area's plans—Karem Sakallah from the University of Michigan (USA) and Ahmed Elmagarmid who, at the time, was at Purdue University (USA). They worked tirelessly to draft detailed strategic documents that were constantly discussed, reviewed, and revised based on input from AES group members in the computing science field. In parallel, Qatar-based institutions conducted surveys related to computing and consulted potential users of select technologies. Sakallah and Elmagarmid also conferred with academic organizations, including Carnegie Mellon University-Qatar and Qatar University, industry leaders like Qatar Petroleum (QP), and relevant Qatar ministries. Ironically, Abdellatif Saoudi, who worked for QP and served as the focal point for

local stakeholders, became the then soon-to-be-formed institute's managing director. He remains in that position today.

Inception of QCRI

In November 2009, Elmagarmid was selected as the inaugural executive director of Qatar Computing Research Institute (QCRI) and subsequently resigned his tenured position at Purdue University. To determine the areas in which QCRI would concentrate, he continued to seek the advice of Sakallah and guided the laborious process that began years earlier with AES and local stakeholders to identify the Institute's key research thrusts. Developing the science and tools needed to advance the Arabic language's standing in the world was one of the Institute's primary objectives. This effort began with projects related to Arabic search engines and handwriting recognition, especially for historical documents. The Institute eventually established four groups: Arabic Language Technologies, Social Computing, Data Analytics (later becoming Artificial

QCRI was created with a mandate to support Qatar's transformation from a carbon economy to a knowledge-based economy.

Intelligence), and Cyber Security (<http://www.hbku.edu.qa/en/qci>).

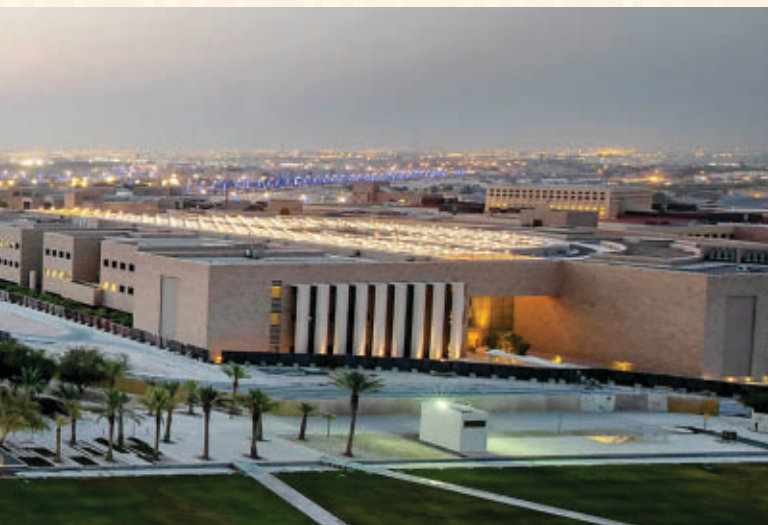
Foundation and challenges. A robust series of discussions that began in 2010 and extended into 2011 helped shape QCRI's four research areas. Management employed study panels for each topic, and renowned computer scientists from around the globe were invited to Doha to participate in roundtable discussions. The panels gathered at least twice for each of the four areas.

The formation of QCRI was not without challenges. The most significant among them was convincing the local community of our relevance and likelihood of success without an existing ecosystem. When the seed for the Institute was planted in 2006, there was no IT industry in Qatar and no computing research to speak of, except for two teaching programs at Qatar University and Carnegie Mellon University-Qatar, both in their nascent stages.

There were other challenges, however. In particular was the recruitment of computer science leaders from top universities and industry, including Carnegie Mellon University, University of California-Berkeley, University of Il-

linois Urbana-Champaign, Purdue University, University of Waterloo, University of Cambridge, University of Oxford, IBM, Microsoft, and Yahoo, among others. In a nutshell, we had to build leaders' trust, ease their concerns about the uncertainties, and ask them to take a leap of faith. Their recruitment had to be handled with care and thoughtfulness. An example of one of our many resourceful approaches was when Yahoo decided to close its research division globally. On the same day of the announcement, we had a team on the phone with their offices in Barcelona, London, California, and New York. We also dispatched a senior scientist to meet with researchers at the Yahoo Barcelona office to explain what we were trying to do. We gained several employees as a result of that visit and continued to mobilize. Hiring the first 10 staff members was an arduous and exhausting process, but we kept one overarching strategy front of mind—identify top people with visibility, recruit and hire them, and then repeat. The emphasis was on track record, pedigree, and reputation.

We also faced the difficult task of convincing top-notch researchers



with tenure that we were focused on longevity. That is, we had the right infrastructure, funding, and leadership to ensure QCRI would exist 10, 25, and 50 years from now. For those who wanted to play it safe and test the waters, we offered them a two-year appointment. Fortunately, most of them stayed, leaving permanent positions in the U.S. and Europe.

Considering QCRI was not a degree-granting institution, attracting research interns and post doctorates proved to be an uphill battle. We had to be innovative in recruiting these positions, opting to form substantive collaborations with as many top-ranked universities as possible and then leveraging those partnerships. We dedicated one-third of our headcount to interns and post doctorates and sought to recruit the best and the brightest annually. This pressure eased as Hamad Bin Khalifa University established graduate programs in computer science, providing us with the opportunity to recruit some of our interns locally.

The challenges were varied, and the solutions



QCRI staff in 2019.

were sometimes unorthodox. The primary takeaway was that one should never compromise on hiring. If you lay the foundation by hiring a smart, creative, and successful team, the rest will fall into place.

Maturation of QCRI

In 2015, QCRI became part of Hamad Bin Khalifa University and continued its mission to help build Qatar's innovation and technological capacity. With a rich mix of academic and industry experts, our 136 scientists and software engineers remain focused

on collaboratively tackling large-scale computing challenges through a multidisciplinary approach. Importantly, QCRI created roles to drive translational academic research and the development of industry-caliber systems.

Guiding the Institute's strategic aims, QCRI also formed an experienced and prominent scientific advisory committee (SAC) that meets twice a year. Today's SAC membership includes: Ruzena Bajcsy, Farnam Jahanian, Hessa Al-Jaber, Yousef Khalidi, Wendy Hall, Michael Wooldridge, Lew Tucker, and Saif Al-Kuwari.

QCRI has matured and is now widely regarded as a formidable organization with highly productive research programs in:

- ▶ **Arabic Language Technologies** focused on machine translation and transcription, natural language processing (NLP), and fake news.

- ▶ **Social Computing** focused on humanitarian action and social good, mobility studies, and automatic persona generation.

- ▶ **Qatar Center for Arti-**

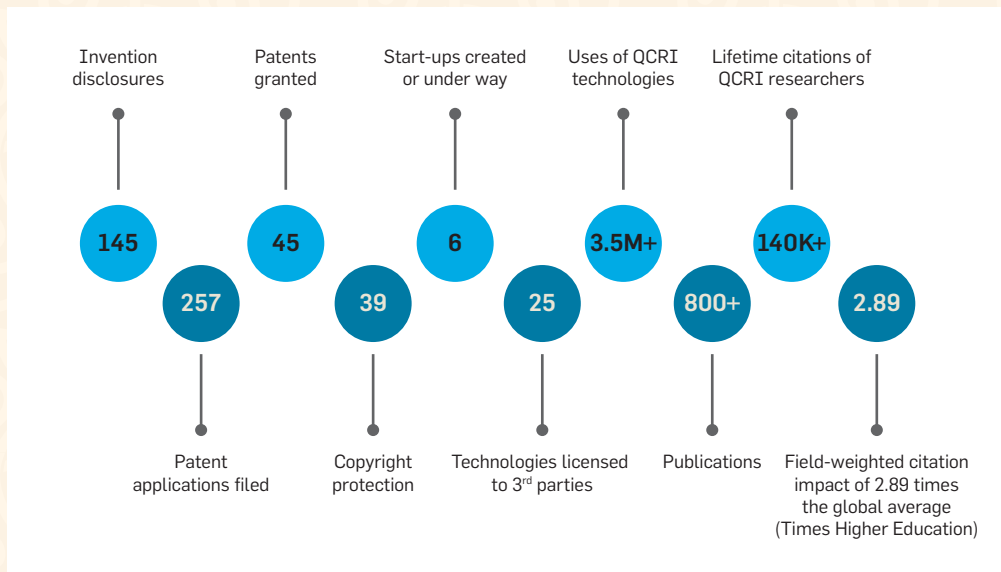
cial Intelligence focused on data discovery and cleaning, transportation and traffic, bioinformatics, and distributed systems.

- ▶ **Cyber Security** focused on serving the needs of government and industry, data analysis, digital forensics, blockchain analytics, and malicious domain detection.

Innovation and Commercialization

Recognizing the importance of translating research in the lab into new or improved products or services in the marketplace, QCRI researchers have created several startups based on our technologies. These include *Tamr* for data cleaning, *Kanari AI* for speech recognition, *Tarantula AI* for 2D- to 3D-video conversion, and *Vorainsight* for automatic persona generation. Our free-access systems have been widely used around the world. This includes *Rayyan*, a tool used for systematic reviews. It currently has approximately 70K active users, most of whom are medical institutions in the U.S. and

Hiring the first 10 staff members was an arduous and exhausting process, but we kept one overarching strategy front of mind—identify top people with visibility, recruit and hire them, and then repeat.



QCRI innovation and commercialization.

Europe. Our *Farasa* system for Arabic NLP also supports millions of Application Program Interface (API) calls monthly, and QCRI's *Shaheen* system has translated more than one billion words between Arabic (with various dialects) and English.

Importantly, our Arabic speech recognition tool is licensed and used within various media outlets, including BBC, DW, and Aljazeera. In addition, QCRI's *QARTA* mapping services have replaced Google Maps in all taxis in Qatar, providing more

accurate routing services through thousands of daily API calls. We have also made a concerted effort to identify needs and develop solutions that promote the widespread adoption of next-generation technical solutions related to humanitarian action and response. As such, QCRI technologies, Artificial Intelligence for Digital Response (AIDR) and *MicroMappers*, have been deployed by the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) during hundreds of

the world's natural disasters and emergencies. In sum, we are keen to ensure the utility of our research is measured by the impact it has on real-world challenges.

Other QCRI technologies have been deployed in local industry and government, including Qatar Airways and the Ministry of Transportation and Communications. Internationally, our technologies have been developed with and used by Boeing, Nokia, and Facebook, among others.

QCRI's contributions are not only documented

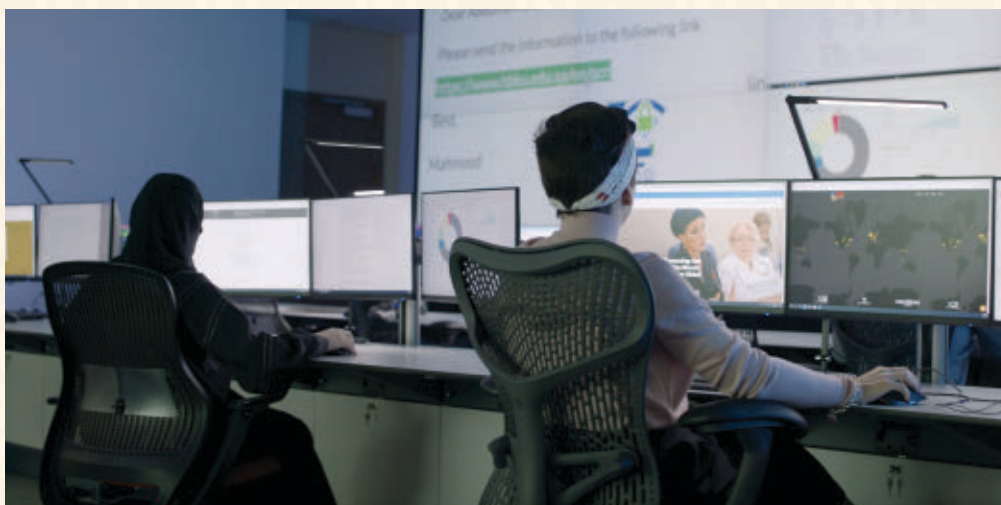
through our tools and systems but also by virtue of our scholarship, professional service, and the many awards, honors, appointments, and boards for which our researchers are frequently chosen (see accompanying figure). As a testament to this productivity, QCRI researchers in the database community actively publish in ACM SIGMOD, VLDB, and IEEE ICDE conferences, with a recent ACM SIGMOD Contribution Award and PC Vice Chair role at ACM SIGMOD and program chair at SIGKDD. In the NLP community, our researchers consistently publish and receive special recognition in ACL and EMNLP conferences. Moreover, we regularly publish in ACM SIGKDD, AAAI, and ICWSM conferences; and have been chosen for the top leadership role in SIGSPATIAL.

QCRI has undergone a metamorphosis. Its transformation from a passing idea to a thriving institute has not been easy, but it has been one of progress. Though filled with ups and downs and sometimes difficult decisions, the journey has also brought great joy originating from a shared commitment to strengthening our community's scientific merits. QCRI's trajectory is bright, and our capacity to make continued meaningful impacts for Qatar, the wider Arab region, and the world is infinite. We invite you all to join us on this journey to recreate the future of the Arab world. 

Ahmed Elmagarmid, QCRI, Hamad Bin Khalifa University, Doha, Qatar.

Abdellatif Saoudi, QCRI, Hamad Bin Khalifa University, Doha, Qatar.

© 2021 ACM 0001-0782/21/4.



The cyber range within QCRI's National Cyber Security Research Lab is where cybersecurity experts are trained to respond to real-world cyberattacks and assess the cyber resilience of digital infrastructure.

Data Science for the Oil and Gas Industry in the Arab Region

BY MOTAZ EL SABAN

OIL AND GAS (O&G) sources will still supply around 50% of the global energy demand by 2040.^a

In this article, we make the case for why the Arab region is well positioned for building world-class data science teams to fill the supply shortage of data professionals,⁵ especially in the O&G field critical to region's economy. This article presents challenges facing O&G industry players, such as governments, regulatory bodies, operators, and investors, and shows how Raisa Energy (with its Egypt-based data science team) is efficiently and effectively solving these challenges. Such challenges aim at assessing the economic viability of an O&G asset that depends on several factors (as shown in the accompanying figure) such as estimating well production, O&G prices, and risks associated with inputs uncertainty. It is worth emphasizing that the challenges presented here are global in nature and yet are tackled with a team fully formed from the region working at a world-class research and development level. We hope this article will motivate aca-



A Zohr gas field in Egypt.

demics and practitioners to tackle challenges within O&G using data science technologies.

The Arab region is well suited for building data science teams serving a global market specially for the O&G industry:

- ▶ There is recent interest from governments in the region to offer data science-related programs and degrees.
- ▶ The region can supply a talented, well-trained workforce at a relatively lower cost.
- ▶ The O&G industry is key in the region; hence data is readily available in

large quantity. Such massive data is the key behind any modern artificial intelligence system.

For example, Raisa Energy, a U.S.-based O&G investment company, has its entire software and data science teams in Egypt building capacity in the important energy domain offering a unique edge for the region. Though junior talent is generally available, there remains a challenge in easily finding senior talent as professionals typically move early into managerial roles for career growth. Our answer to this chal-

lenge is twofold: create opportunities for juniors to grow technically by working on challenging problems of global nature, and complement juniors by experienced returning expats to the region.

We next detail some of the technical challenges that Raisa Energy faces and the novel approaches it uses in solving them that resulted in several academic publications and U.S. patents.¹⁻⁴

Well production forecasting is a time series forecasting problem of an O&G well production. Well features include geological

a <https://www.brief.com.cy/sites/default/files/2019-10/BP%20Energy%20Outlook%202019.pdf>

maps, engineering parameters used in drilling and completing the wells such as the amount of water and chemicals, location information, well lateral length, and production of neighboring wells. Raisa introduced a state-of-the-art production forecasting system using an ensemble of a random forest and a sequential deep learning model trained on 50,000 wells^{1,2} and has been granted a U.S. patent on the topic. Raisa's approach is the first of its kind to utilize a comprehensive set of features, leverage data in the range of thousands of wells and use advanced machine learning methods through experimentation. Our models are tested on independent test sets both offline (on older held-out sets) and online in the day-to-day job of reservoir engineers. When measured on held-out sets our models achieve close to 90% prediction accuracy. To cope with the ever-growing nature of well data, models are continuously retrained using our in-house continuous train-

ing platform. Whenever a retrained model scores higher than existing models on test sets it is pushed to production.

In addition to estimating a single-point estimate production curve, it is important to estimate a probabilistic forecast to model variability in input features. We estimate different percentile curves by adapting machine learning loss function to include quantile loss.³ Finally, we use production forecasting models to answer what-if questions such as what happens if we increase a chemical? Such capability is crucial for optimizing well operation for instance.

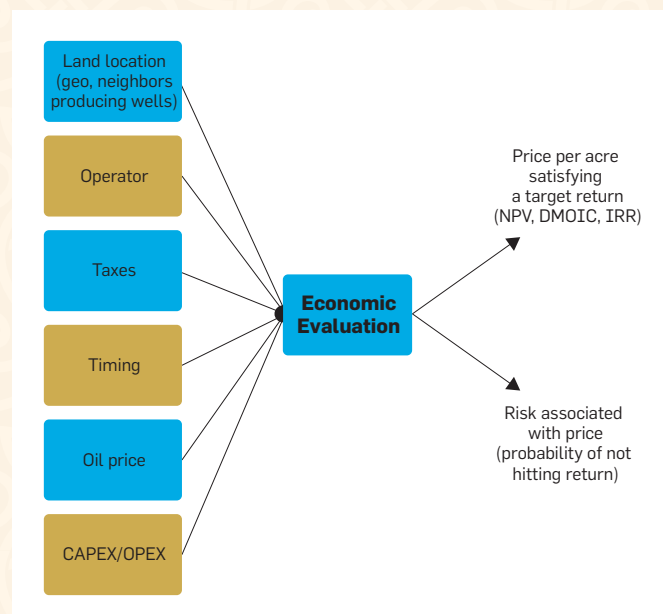
Timing indicators for well life cycle events. Investors in O&G need to estimate when well production begins, which signals the start of cash flow. An investor does not control, though the drilling schedule of lands owned by an operator and hence machine learning can be used to reverse engineer operator logic in selecting lands to be drilled next.

The O&G industry is a risky business as there is large uncertainty in input factors values such as geology, oil prices, and the effect of neighboring wells.

This problem is unique to Raisa as an investor in the O&G industry and there is very little work done on this technically challenging problem. Raisa has been making strides tackling this problem, which led to a pending U.S. patent. The main idea of our approach is to use a comprehensive set of features to model operator rig movement from location to location and hinge on the machine learning model to optimize mapping from these features to the selected next best locations to move. Features include production volume estimation at target site, distance to pipeline infrastructure, automatic detection of drilling pads using satellite imagery,² and estimated oil price to model operator movement using publicly available historical data gathered by GPS-powered drilling rigs. For example, we have researched and implemented a deep learning image segmentation method based on U-nets to detect drilling pads, which are clear indicators of imminent drilling at the site.² Our work on semantic image segmentation was among the top four competing methods in the INRIA challenge.²

Operator movement from location to location is then modeled as a sequential decision problem and our current solution relies on optimizing a step-by-step classifier to build the optimal operator trajectory. Once we build this classifier, we test various operator models using a subset of the already drilled locations achieving around 80% classification accuracy of next location prediction for operators with enough historical samples. We have also experimented with more elaborate approaches such as reinforcement learning by modeling the problem as a Markov decision process (MDP). Our initial experimental results are very encouraging.

Risk analysis. The O&G industry is a risky business (whether in exploration or production) as there is large uncertainty in input factors values such as geology, oil prices, and the effect of neighboring wells. To mitigate such risks, several scenarios governing input parameters are studied, along with correlation between inputs. For example, we estimate probability distributions for well production, geological parameters, and oil price estimates and



Economic evaluation of oil and gas assets.

hence it is of paramount importance to have probabilistic predictions on key variables of interest. We also estimate correlations among these variables. The goal is to build a multivariate probability distribution of inputs with the dollar value of the O&G asset being the output. We use Monte Carlo-based methods to sample (typically 2,000 samples are used) from input distributions and then run the samples through our economic model to estimate output cash flows.

With such an output distribution at hand, one can estimate the level of risk associated with an O&G asset. It is important to note that Monte Carlo simulations are computationally expensive and hence usually require fast sampling strategies (such as lattice hypercube sampling) and use of parallelization or distributed processing. Another closely related challenge to risk estimation is to find an optimal risk-reward portfolio of assets owned by a single investment company for instance. The goal is to build a risk-reward profile of all assets owned by Raisa and then decide on new opportunities based on its economic

rvalue as well as where it fits with already owned assets in terms of risk-reward profile. Through careful modeling of input distributions and a highly optimized parallelized implementation, Raisa's risk simulation solution is on par with expensive simulation software giving Raisa a unique competitive edge in the market.

Document information extraction. Much of O&G data comes in the form of text containing well cost information, production reports, and legal documents describing drilling restrictions. Automatic information extraction technology can be a huge time saver in these scenarios. For example, at Raisa we automatically extract cost information from well documents using semi conditional random fields (CRF) models to achieve highly accurate results (typically more than 85% F1 on fields of interest) on a challenging problem with little labelled data and high variability in document structure and templates. Over the years, we have built a full text processing pipeline that handles various document formats including those requiring OCR. Typical processing blocks include

Raisa's risk simulation solution is on par with expensive simulation software giving Raisa a unique competitive edge in the market.



An oil and gas processing plant in Egypt.

page layout extraction, table extraction and slot filling. Once information is automatically extracted it can be stored in databases enabling fast search and retrieval.


Other Challenges

Beside the challenges mentioned here there are other technical challenges in the O&G field where data science can provide effective solutions. We list a sample here for the sake of completeness.

► **Oil price estimation.** Oil price is key in estimating cash flow generated by an O&G asset. This is a time series forecasting problem where deep learning methods have been used such as Gated Recurrent Units (GRU) architectures. In other work, researchers analyzed news articles to estimate oil price movement using deep networks.

► **Model interpretability.** O&G decision makers seek transparent ML models enabling them to understand the effect of different features on prediction. Hence it is crucial to couple accurate, and often

complex, prediction models with model interpretability approaches.

► **Regulating oil prices.** Regulatory bodies such as OPEC monitor production of member countries to control supply. One recent use of machine learning in this area is through automatic image understanding of oil tanks filling levels from satellite imagery. 

References

1. Amr, S., El Ashhab, H., El Saban, M., Caile, C., Schietinger, P., Kaheel, A. and Rodriguez, L. A large-scale study for a multi-basin machine learning model predicting horizontal well production. In *Proceedings of the SPE Annual Tech. Conf. and Exhibition*, (Dallas, TX, USA, Sept. 24–26, 2018).
2. Khalil, A. et al. Large-scale semantic classification: outcome of the first year of INRIA aerial image labelling benchmark. IGARSS 2018.
3. Mostafa, H., El Mahdy, M., El Saban, M. and Abo El Kheir, M. Probabilistic time series forecasting for unconventional oil and gas producing wells. In *Proceedings of the IEEE Conf. Novel Intelligent and Leading Emerging Sciences*, 2000.
4. Systems and methods for optimizing production of unconventional horizontal wells, United States Patent Application 20190284910.
5. State of AI Report, 2020; <https://www.stateof.ai/>

Motaz El Saban is Director of Data Science of Raisa Energy, and associate professor in the Faculty of Computers and Artificial Intelligence at Cairo University, Egypt.

The Strategic Pursuit of Artificial Intelligence in the United Arab Emirates

BY FARAH E. SHAMOUT AND DANA ABU ALI

ARTIFICIAL INTELLIGENCE (AI) is expected to mark a paradigm shift in the fabric of our society. To harness the societal and economic benefits of AI and be best prepared for future challenges, it is more important than ever to be strategic about investing in AI. As such, the pursuit of AI in the United Arab Emirates (UAE) has been unique and visionary in recent years. Here, we provide a high-level overview of key efforts that are significantly contributing to the UAE's strategic pursuit of AI, namely the national vision and strategy, research infrastructure, capacity-building, AI adoption, and cross-sector collaborations.

The UAE government's vision is to make the UAE a world leader in AI by 2031, as per the National Artificial Intelligence Strategy launched in 2017.¹ The strategy's objectives include positioning the UAE as a central AI hub in the region and globally, developing capabilities and local talent, and adopting AI in both public and private services to boost performance. The launch of this strategy also coincided with the appointment of the UAE's, and the world's, first Minister of State for Artificial Intelligence HE Omar Sultan Al Olama and the launch of



Attendees at the first AI Everything Summit in Dubai on May 1, 2019.

the Emirates Council for Artificial intelligence and Digital Transformation. Several initiatives have also emerged under the leadership of the National Program for AI, which builds and shares resources in the pursuit of the UAE's policy objective to be a global participant in the responsible use of AI²—the motto of the National AI Program is “B.R.A.I.N: Building a Responsible AI Nation.” This includes annual AI Everything Summit for Governments and Businesses, the AI Code Hub platform that hosts open source software developed locally in the UAE,⁴ and an AI retreat that attracted more than 350 AI experts from the public and private sectors.² As such, the government's visionary approach to investing in AI

has become a key driver to AI developments across all sectors in the country.

Research Infrastructure

Investment in a strong research infrastructure is one of the key endeavors of the UAE's AI pursuit. According to the AI Hardware Infrastructure Report that was published in November 2020,⁴ the UAE has “the 36th most powerful high-performance computer in the world” according to the Top500 list for November 2020. This reflects a heightened interest in enabling research and development efforts through investing in computational hardware infrastructure. While 89% of the total processing power is dominated by the private sector, academic

institutions are also home to state-of-the-art high-performance computing (HPC) systems. For example, at New York University Abu Dhabi, the HPC system, also known as “Dalma,” supports research in computer vision, health informatics, natural language processing, ocean and climate modeling, computational astrophysics, and bioinformatics. The United Arab Emirates University also owns three HPC systems to support a variety of scientific research projects. Increased research activity in AI in both academia and industry will have a long-term impact on the region's global competitiveness in knowledge creation, innovation, and talent development.

Building Talent and Capacity

To develop capabilities and talent in AI as part of the strategy's objectives, several learning development initiatives have been implemented to target different audiences and age groups. For example, in partnership with Kellogg College at the University of Oxford, the National AI Program developed an executive-level coursework program to train UAE nationals in key government positions to accelerate the delivery of the national strategy.² The UAE AI Camp was also the first in the region

The UAE government's vision is to make the territory a world leader in artificial intelligence by 2031.

to offer spring and summer camps for high school and university students. Most recently, the Mohamed Bin Zayed University of Artificial Intelligence was launched as a graduate-level institution offering MSc and Ph.D. programs specialized in AI. It is expecting its first cohort to arrive in 2021.⁵ Such opportunities will lead to informed use and adoption of AI on a national scale and will also further attract global talent through the competitive graduate programs.

Adoption of AI

The strategic national AI vision in the UAE has driven entities in the public and private sector alike toward the adoption of AI. A Microsoft AI report conducted in 2019 found that UAE organizations show a “lead in proactiveness in adopting AI solutions, when compared with global peers. Some 70% of double-digit growth companies in the UAE intend using AI within the coming year to improve decision-making, as opposed to 46% worldwide.”⁶

The UAE healthcare sector is a good example of how government entities are striving to capitalize on these opportunities. The Ministry of Health and Prevention developed and released Medopad, a smart application that continuously monitors and analyzes patient data to predict life-threatening medical condi-

tions and allow for proactive healthcare services.⁷ Dubai Health Authority and Agfa HealthCare have partnered together to utilize AI-enabled workflows in medical imaging to enhance the diagnosis of TB.⁸ Robotics also presents a promising opportunity in the sector, presented in the adoption of robotic surgeries and pharmacies.⁷

Similarly, other government entities launched new AI pilots by collaborating with established and rising technology companies to augment their services using AI. Rashid, Dubai Smart Government's call center virtual agent, offers official answers to customers' questions about procedures, documents, and requirements needed to conduct various transactions in Dubai. The virtual assistant was developed in the Dubai AI-Lab, in cooperation with the Smart Dubai Office and IBM, using WatsonAI. Dubai's Road and Transportation Authority has been following a similar trajectory, gradually integrating AI into its services, with autonomous taxis—the first in the region—being one of their top projects.⁹ In October 2020, the authority announced a new data strategy, aligned with the UAE 2031 AI strategy, to enhance efficacy and reduce costs using AI. Those initiatives illustrate a strong public commitment to meet the objectives of the

strategy and to enhance efficiency of public services.

A Collaborative Ecosystem

The strategic synergy between academia and research, private sector companies, and government agencies has produced a plethora of promising use cases that cannot be covered fairly within the scope of this article. AI companies in the private sector have had a significant impact on the progression of AI in the country through the use of their own proprietary research and innovation. In November 2019, the UAE launched an initiative under the name “UAE AI Network,” with the goal of bringing together public and private sector organizations and academic institutions under the umbrella of helping achieve the nation's AI strategy.² This is crucial for harnessing the best ideas in the development of AI and to achieve societal impact.

Future Outlook

The UAE has become a central AI hub under the guidance of its national strategy. The aforementioned initiatives are only examples of numerous existing and emerging efforts. According to research conducted by the IDC, the spending by federal/central governments on AI across MEA is projected to see the strongest growth of the region's top five verticals, increasing at a compound annual growth rate of 26.3%.¹⁰ This investment, as what is happening in the UAE, is likely to continue to lead to significant development of local talent in the region and increase in innovation and the quantity and quality of research and development projects, both in industry and academia. To leverage the advantages

of such a fast-paced field, various stakeholders must continue to aim for constant learning, talent development, and reinvention. In the next decade, AI will begin to play a pivotal role in our daily activities across several domains, such as healthcare, transport, and education, to name a few. The prioritization of the responsible use of AI as a guiding principle of the UAE AI strategy will ensure the adoption of AI will serve the betterment of humanity. 

Additional Resources

1. UAE Strategy for Artificial Intelligence. The Official Portal of the UAE Government; u.ae/en/about-the-uae/strategies-initiatives-and-awards/federal-governments-strategies-and-plans/uae-strategy-for-artificial-intelligence.
2. National Program for Artificial Intelligence Official Website; <https://ai.gov.ae>.
3. AI Code Hub Github; <https://github.com/artificial-intelligence-office>.
4. AI Hardware Infrastructure Report UAE. National Program for Artificial Intelligence; https://ai.gov.ae/wp-content/uploads/resources/AI_Hardware_Infrastructure_Report_UAE_2020_EN.pdf.
5. Mohamed bin Zayed University (MBZUAI) Official Website; <https://mbzuai.ac.ae/>
6. High growth companies in the UAE ready for AI adoption: Microsoft AI report. Microsoft News Center Middle East & Africa; <https://news.microsoft.com/en-xm/2019/04/02/high-growth-companies-in-the-uae-ready-for-ai-adoption-microsoft-ai-report/>
7. Robotics and AI applications. The United Arab Emirates Government Portal; <https://u.ae/en/about-the-uae/digital-uae/robotics-and-ai-applications>.
8. Late-breaking abstract: Use of AI in accurate diagnosis of TB using medical Imaging: A promising result from UAE. *European Respiratory J.* 52, 62 (2018); https://erj.ersjournals.com/content/52/suppl_62/OA5171.
9. Building an AI nation: Accelerating artificial intelligence adoption through agile policymaking—The case of the UAE. *Dubai Policy Review*; <https://dubaipolicyreview.ae/building-an-ai-nation-accelerating-artificial-intelligence-adoption-through-agile-policymaking-the-case-of-the-uae/>.
10. Annual Spending on Artificial Intelligence in the Middle East & Africa to Top \$530 Million by 2022. *International Data Corporation*; <https://www.idc.com/getdoc.jsp?containerId=prMETA45065619>.

Farah E. Shamout is Assistant Professor Emerging Scholar at New York University Abu Dhabi.

Dana Abu Ali is a software engineer at MEA Clients Center at IBM Dubai.

An AI-Enabled Future for Qatar and the Region

BY ASHRAF ABOULNAGA, SANJAY CHAWLA, AHMED ELMAGARMID, MOHAMMED AL-MANNAI, AND HASSAN AL-SAYED

QATAR IS A small peninsular nation on the northeastern coast of the Arabian Peninsula. Qatar is endowed with abundant hydrocarbon resources and is the world's largest producer of liquified natural gas (LNG), which accounts for over 80% of its export earnings. Like many of its wealthy neighbors, Qatar faces a unique dilemma with the onset of artificial intelligence (AI) technologies. Despite having one of the world's highest per-capita income and a highly educated local population, the majority of Qataris are under-employed and working in government white collar jobs where they are unable to fully realize the potential of their level of education. These are precisely the occupations that are likely to be made redundant by AI.¹ The bulk of the workforce in Qatar consists of expatriates drawn primarily from



South Asia and the Middle East and North Africa (MENA) region. As the finite horizon of a natural resource-based economy comes closer, countries like Qatar have no option but to embrace AI to transition into a knowledge-based economy while protecting and perhaps enhancing their current standard of living.

In October 2019, and in

collaboration with Qatar Computing Research Institute (QCRI), the government of Qatar released a National Strategy for AI.³ The aim of the strategy is to provide decision-makers and the wider public in Qatar with a nuanced and realistic view of AI technology and at the same time serve as a “call for action” toward a future where AI will become the defining technology of the 21st century and beyond.

Pillars of AI

The strategy is divided into sections (referred as pillars): Race for Talent, AI-Augmented Jobs, Knowledge Economy, Data and Computing Infrastructure, Ethics and Governance of AI, and AI+X Future (as depicted in the figure here). For each of these pillars,

the strategy makes Qatar-specific recommendations in order for the country to realize its national vision to transition into a knowledge-based economy by 2030. Here, we briefly outline and contextualize the pillars and report on progress since the strategy was announced.

Race for talent. Some 95% of the workforce in Qatar consists of non-Qatari expatriates. A large bulk of the workforce consists of blue-collar workers employed in the construction and service sector. As the infrastructure work related to the FIFA 2022 World Cup comes to an end, Qatar must create pathways to attract talent that has experience working in AI and digital ecosystems. For example, fast-track visas for AI engineers can help build a critical mass of AI

In Qatar, the trade-off between AI and job loss will not be that stark because the majority of the Qatari population is overeducated and underemployed.

talent in the country—at least in the short term. The AI talent can be deployed in existing strategic industries ranging from oil and gas, banking, utilities, and telecommunications. For a more sustainable and long-term solution to create AI talent, the strategy proposes a measured transformation of the curriculum at all levels of education. Modern AI rests on the ability to transform data and students must be taught early on how to work with data and develop a data-driven approach to problem solving, analysis, and critical thinking. For example, different types of data associated with COVID-19 could be used in a mathematics class to introduce “curve-fitting” or in a biology class to introduce

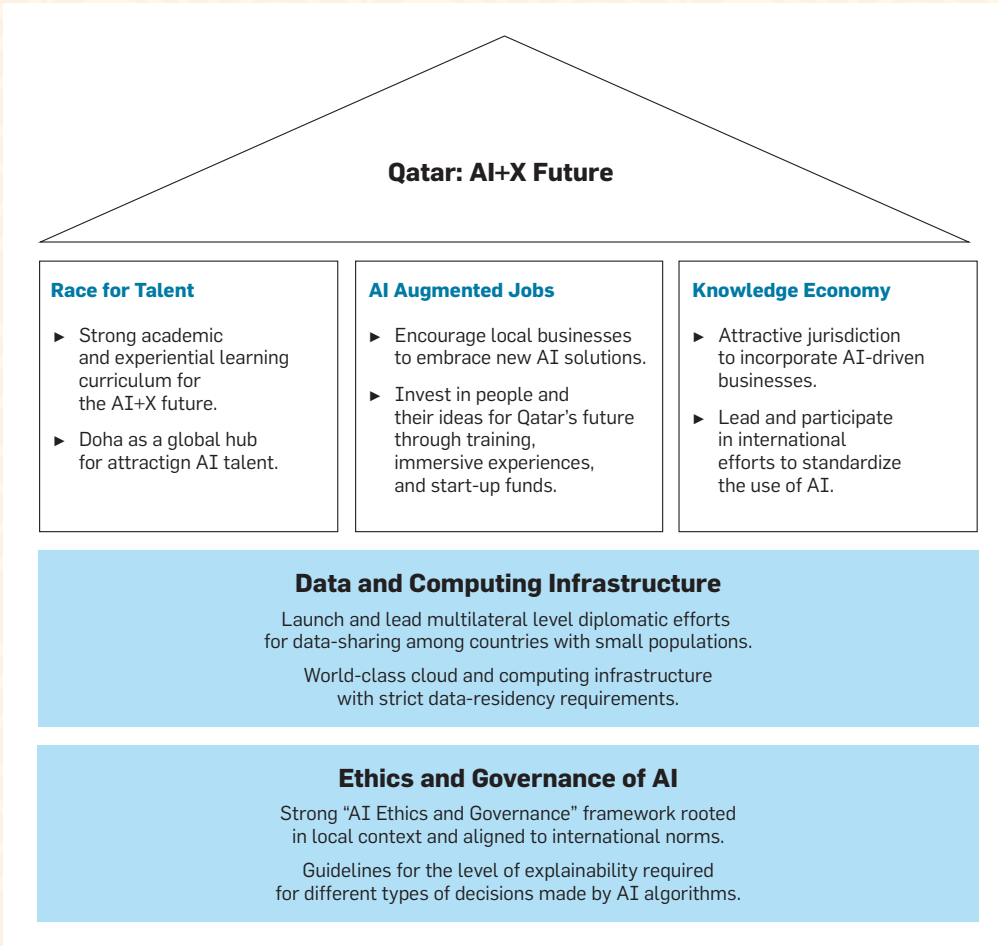
protein sequence structure or in a geography class to compare the spread of the virus across different parts of the world.

AI-Augmented Jobs. A recent study by QCRI, as a follow up work on the strategy, indicates a massive impact of AI in the Qatari workplace.² Nearly 45% of tasks associated with jobs in which Qataris are employed can be augmented with existing commercially available AI technology or for which a patent has been approved. Similarly, a large number of occupations where non-Qataris are employed will be impacted. Nearly all studies on the impact of AI on the workplace indicate that white collar jobs are more susceptible to be replaced by AI than blue collar jobs. Since a

majority of the Qatari population is engaged in white collar occupations, they are likely to be disproportionately impacted. Policy makers around the world are facing the dilemma of embracing AI technology and at the same time preventing massive job losses that could lead to social and political upheaval. In Qatar, the trade-off between AI and job loss will not be that stark because the majority of the Qatari population is overeducated and underemployed. Creating training programs that can help them transition to an AI-augmented work environment will be relatively easier in Qatar than in other parts of the world. Qatar has already built a world-class infrastructure in education and

research and many major international universities have a branch campus in the country. For example, Carnegie Mellon University, which arguably is the fountainhead of AI research and education, has a branch campus in Qatar where undergraduate degrees in computer science and information systems are offered. CMU’s expertise in AI could be leveraged to create training courses to help transition workers to an AI-augmented workplace.

Knowledge Economy. AI can serve as a catalyst for Qatar to transform itself into a knowledge economy as envisioned in the Qatar National Vision 2030 plan. Qatar needs to transition from an economy based on hydrocarbon and downstream industries to an economy grounded in data and AI. An ecosystem should emerge where talent, ideas, and investment are allowed to interact in a harmonious manner. Qatar’s investment in research centers like QCRI can be the genesis of such an ecosystem. QCRI has already developed specialized AI technology in Arabic speech translation, urban computing, fake news detection, data integration, persona generation, and systematic reviews and these technologies have already been spun-out into start-ups. Processes have been put in place where researchers can go on “entrepreneurial leave” to transition these start-ups to their next level of maturity. The next step is to update legal frameworks where investors can enter and exit the market in a transparent manner. Local talent must be harnessed into the ecosystem to ensure long term sustainability.



The pillars of AI.

Data and Computing Infrastructure: Modern AI is based on data and requires a robust computing infrastructure. Qatar should take a leading role in multilateral efforts for data exchange as it had done in the World Trade Organization (WTO) by hosting the Doha round in 2001. Several barriers against the free flow of data are emerging around the world as concerns of data privacy and security are leading countries to inhibit the flow of data. While the strategy was written before the pandemic, it foreshadowed the importance of data exchange. For example, WHO initiated Solidarity Trials for finding effective treatment for COVID-19 involved over twelve thousand patients in five hundred hospitals across thirty countries. Data on treatments was exchanged and collected by the WHO and several studies have been published for evaluating COVID-19 therapeutics.⁵ In the 2019 G20 summit in Osaka, an initiative Data Free Flow with Trust (DFFT) was started to create standards and policies encouraging a seamless flow of relevant data across the globe while ensuring safeguards arounds privacy and security.⁶ Qatar should form a coalition among small nations to ensure that they have an effective voice in global forums on data exchange.

Ethics and Governance of AI. While AI models are extremely accurate and outperform humans on many benchmark cognitive tasks, their decision-making logic often remains opaque and uninterpretable. As AI models are based on data, they may inadvertently capture and amplify social biases that are inconsistent

with the laws of the land. For example, in the U.S., AI models for determining the length of a jail sentence for a convicted criminal amplified racial biases even though racial information was not provided to the model.⁴ Similarly social media platforms have “hard coded” maximizing retention as their objective and will recommend content to a user which often leads to extreme polarization in a society. As AI technology permeates into important sectors of society like education, health and law, care must be taken that recommendations from AI systems are ethical and remain consistent with local ethical norms and practices.

An AI+X Future. As a small country with a limited technological base and talent, Qatar will remain a net importer of AI technology for the foreseeable future. However, the strategy has identified a few niche domains where Qatar can be an important player in the world stage. For example, in precision medicine, new AI-driven technology can be built around the data emerging from the Qatar Genome Project (QGP). In particular Qatar can take the lead in developing specialized AI tools to detect, preempt and manage diseases associated with the practice of consanguinity. In the domain of digitalization of the Arabic language, Qatar-based international news media, academic entities, and research institutes can combine forces and develop novel AI technology. Qatar has invested billions of dollars in sporting infrastructure and can encapsulate the learning in AI tools which can be exported to international

Qatar needs to transition from an economy based on hydrocarbon and downstream industries to an economy grounded in data and AI.

markets. As the world’s largest producer and supplier of LNG, Qatar can use the vast amount of associated data generated from the gas fields to create niche AI products for the hydrocarbon industry. Finally, in the area of transportation, where Qatar has invested huge amounts of resources to become a “smart nation,” the triangulation of data, talent and investment can create cutting-edge AI technology.

Strategy Update

The strategy was released in October 2019 and by March 2020, Qatar like the rest of the world, was in the midst of an unprecedented pandemic. As the world moved online, the role of data and AI has become even more salient than before. From the outset, the WHO warned that the world was not only witnessing a pandemic but an “infodemic” where unreliable and false information was interrupting efforts to control the pandemic. Thus, the importance of “Ethical AI and Data” has clearly stood out during the pandemic. On the positive side, the democratization of COVID-19 data ensured local AI models could be developed to forecast the spread of the disease. In fact, data-driven models developed by QCRI to forecast the epidemic curve and mobility patterns were

extensively used by policy makers in the country.

The government of Qatar is in the process of forming an implementation and oversight council to streamline AI activities based on recommendations of the strategy. The council will have representation from all arms of the government, academic and research institutions and civil society. The brave new world of AI is taking root in Qatar and the region. 

Additional Resources

1. Frey, C.B. *The Technology Trap*. Princeton Press, 2019.
2. Qatar Center for AI. November impact of AI on Qatar’s labor market, (2020); https://qcai.qcri.org/wp-content/uploads/2020/11/Impact-of-AI-on-Qatars-Labor-Market_15Nov2019.pdf
3. Qatar Computing Research Institute. National AI Strategy for Qatar, (2019); <https://qcai.qcri.org/wp-content/uploads/2020/04/QCRI-Artificial-Intelligence-Strategy-2019-ENG.pdf>
4. Rudin, C., Wang, C., and Coker, B. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review* 2, 1 (2020). <https://doi.org/10.1162/99608f92.6ed64b30>
5. WHO. Solidarity Trials, (2020); <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments>
6. World Economic Forum. Data Free Flow with Trust; http://www3.weforum.org/docs/WEF_Paths_Towards_Free_and_Trusted_Data%20_Flows_2020.pdf

Ashraf Aboulnaga, Qatar Computing Research Institute.

Sanjay Chawla, Qatar Computing Research Institute.

Ahmed Elmagarmid, Qatar Computing Research Institute.

Mohammed Al-Mannai, Ministry of Transport and Communications.

Hassan Al-Sayed, Ministry of Transport and Communications.

© 2021 ACM 0001-0782/21/4.

Entrepreneurship Ecosystem in Lebanon

BY WALID R. TOUMA AND SAAD EL ZEIN

IN LEBANON, AND way before the arrival of seed money, venture capital, and other forms of equity support for start-ups and innovative ventures, the main source of support came from the SMEs bank loans guarantee programs provided by Kafalat.^a With successful platforms implemented by the Lebanese Central Bank through initiatives such as Kafalat, the entrepreneurship ecosystem started to take shape and flourish through universities in the early 2000s. The catalyst was support from the Lebanese Central Bank in 2014 in the form of Circular 331,^{b,c} inject-

a <https://www.banqueduliban.gov.lb/files/tabs/Kafalat.pdf>

b <http://2015.bdlaccelerate.com/everything-you-need-to-know-about-bdl-circular-331/>

c <https://bd1.gov.lb/circulars/intermediary/5/37/0/Intermediate-Circulars.html>



ing an estimated US\$400 million in the Lebanese enterprise market through private equity funds, with the support of local banks. Such support gave rise to the U.K. Lebanon Tech Hub in 2015 (UKTH), an international accelerator initiative between the U.K. government through

the U.K. embassy and the Lebanese Central Bank, in addition to the launch of new accelerators and incubators including Speed in 2015, Flat6Labs, and Smart ESA in 2016. In parallel, the World Bank supported the launch of the iSME Fund, a private equity fund entrusted to Kafalat to manage in 2015, and Circular 331 provided the support to launch Berytech Funds, Broadgate Y Venture Partners, Middle East Venture Partners, and many more venture capital and private equity funds.^d

Success Stories

Despite the challenges that Lebanon has faced over the past five years, several suc-

cess stories materialized in its entrepreneurial ecosystem. Since 2015, UKTH/The Nucleus Ventures accelerated 100 start-ups, created 2,000 jobs, with 42 start-ups reaching the seed funding stage, 10 series A, four series B stages, and one successful entry in the online gaming space. Three current companies in the Nucleus Ventures portfolio have valuations above US\$10 million, with Proximie valued at over US\$30 million.

Speed, another solid accelerator in the Beirut Digital District, accelerated 42 start-ups through six cycles, with one of its venture companies, NAR Technologies, getting acquired by U.S.-based (San Antonio, TX) B3Bar Holdings. NAR Technologies was founded by two Lebanese American

The ecosystem of the entrepreneur in Lebanon will need to adapt to a different form of deliverables management, along with the newly burgeoning support of academic institutions.

d The Lebanese venture capital and private equity funds launched their own association in 2018; <http://bit.ly/3mEYHot>

University (LAU) graduates—Charlie El Khoury and Nicolas Zaatar. NAR identified the opportunity in dramatically improving the labor-intensive and time-consuming process of data analysis and reporting performed following each inspection flight by drones.

Kafalat, through the “Grant” component of the iSME Fund, deployed the full US\$2.5 million by the end of 2019. The fund reached 175 beneficiaries with grants of up to US\$15,000. Over 65 of those beneficiaries have been able to attract investments of over US\$15 million from venture capitalists and investors to date, and numbers keep on rising. Moreover, through the “co-equity” component of the iSME Fund, Kafalat deployed US\$14.5 million up till 2019 in 21 start-ups. The co-investment in these companies from the local venture capital funds had reached US\$38.8 million for the same dates. On the bank loan guarantee front, out of the 41 Kafalat INNOVATION loan guarantees, 23 loans were repaid and their funded companies continue to thrive, achieving a 56% success rate in such high-risk invest-

ments. Moreover, out of 91 Kafalat PLUS loan guarantees to technology companies, 79 loans were repaid, achieving a super-stellar 86.8% success rate for such start-up loans.

The three most active universities in the Lebanese venture development and acceleration processes are USJ, USEK, and LAU. USJ Berytech and its funds accelerated over 130 companies, with over 3,600 entrepreneurs supported, US\$70+ million invested in start-ups, and four major exits. Currently, Berytech has 31 companies in its portfolio, and supports a large number of SMEs through the USAID-LED program. USEK Acher Center accelerated 25 start-ups in one cycle, with more than US\$4 million raised for its portfolio in a very short period of time. Last, LAU FMIC accelerated nine start-ups in one cycle, with limited amounts of funding raised to reach the prototype stage by five of the nine start-ups.

More to Be Done

As with all major initiatives, there is always more to be done, and Lebanon is no exception. The

An overhaul of the Lebanese commerce laws is long overdue, and both fund managers and entrepreneurs should take the lead.

Central Bank’s Circular 331 restricted the capital support to only Lebanese companies; with such export-driven ecosystem, the need by locally funded companies to scale up outside Lebanon to facilitate trade and local services to international clients got chocked up by the lack of international funding that Circular 331 did not support. An updated Central Bank circular in this regard should be requested by the stakeholders of this ecosystem.

In addition to the much-needed private equity risk management expertise, the Lebanese commercial laws have not been developed enough to meet the needs of the structured approach with private equity investments. In particular, the Lebanese commerce laws lacked any legal premise for a Lebanese corporation to issue stock options, stock warrants, and shares of stock with preferred rights (preferred stock). Accordingly, an overhaul of the Lebanese commerce laws is long overdue, and both fund managers and entrepreneurs ought to take the lead.

Hope

The ecosystem of the entrepreneur in Lebanon

will need to adapt to a different form of deliverables management, along with the newly burgeoning support of academic institutions. By leveraging Lebanon’s human capital, leading universities such as LAU, AUB, USJ, ESA, and USEK will play key roles in providing the critically needed support to nurse the new entrepreneurial initiatives in Lebanon by providing both space and coaching to the emerging new entrepreneurial ventures. Financial support will be driven by both local and international funding structures, all hinging on sound business concepts and the solid execution by the local teams leading such ventures. Combining these strengths will provide the oxygen needed to push forward an innovation-driven economy in Lebanon, an economy in critical need of hope, perseverance, and above all, the solidarity behind delivering a stellar performance in such a dynamic ecosystem. 

Walid R. Touma is Director of the University Enterprise Office at Lebanese American University, Beirut.

Saad El Zein is head of the LAU Fouad Makhzoumi Innovation Center at Lebanese American University, Beirut.

© 2021 ACM 0001-0782/21/4.



Autonomous Driving in the Face of Unconventional Odds

BY HESHAM M. ERAQI AND IBRAHIM SOBH

TRAFFIC ACCIDENTS ARE a major unsolved problem worldwide. Yearly, it causes around 1.35 million deaths and 10 million people sustain non-fatal injuries⁹ in addition to having substantial negative economic and social effects. With approximately 90% of accidents being due to human errors, autonomous driving (AD) will play a vital role in saving human lives and substantial property damage. Moreover, it promises far greater mobility, energy saving, and less air pollution.

Despite the recent advances to achieve such promising vision, enabling autonomous vehicles in complex environments is still decades away.⁶ The problem turned out to be more difficult than expected and it is even harder in the extremely complex and challenging driving environments in many regions around the world, including most of the Arab region. In

the coming subsections, we discuss the challenges facing the successful implementation of AD in the Arab region and map them to the four pillars of the Autonomous Vehicles Readiness Index.⁷ Finding solutions to those challenges will help deliver the benefits of AD to world regions that are in desperate need for it, as more than 90% of traffic deaths occur in low- and middle-income countries with Africa having the highest death rates.⁹ In addition, it is very beneficial to the automotive industry as 24% of the global automotive market sales are outside Europe, China, U.S., India, and Japan.⁵

Infrastructure challenges.

The AD algorithmic pipeline employs a ‘sense-plan-act’ design,³ which is the basis of many robotic systems. Advanced sensors allowed for a more accurate sensing of the environment and the surrounding objects. Nevertheless, today’s road infrastructure provides many open challenges to be fully solved in ‘planning’



the vehicle’s actions based on understating the driving scene and eventually ‘acting’ by commanding the vehicle’s control system. Based on our literature review, we conclude that road infrastructure quality is determined by 10 features: signs, marking, barriers, lane, shoulder, median, right-of-way, horizontal alignment, vertical alignment, and lighting. The number of features contributing to road safety problems increases considerably in the case of developing countries and the Arab region⁴ compared to developed countries. Arab-region specific problems include a combination of lack of lane markings, traffic signs, light poles, and roadside barriers. Add to that the higher traffic densities and unofficial pedestrian waiting areas, captured in Figure 1, and it creates a very challenging driving scene for the ‘plan

and act’ algorithms. From another perspective, once adopted, a self-driving car is expected to generate more than four terabytes of data daily, while the communication infrastructure in many parts of the Arab region is not ready to accommodate such traffic requirements.

The solution to improve road infrastructure should start by deploying Road Asset Management Systems (RAMS) to structurally plan and implement maintenance. In low-income countries, RAMS data collection process should be more frequent due to the low-cost material used, and hence it is more costly. Moreover, there is a lack of qualified experts for data analysis. One solution to the cost problem is to adopt more pervasive solutions encouraging drivers to participate in data collection and to rely on the modern deep learning-

According to a recent survey, approximately half of UAE residents are likely to own a self-driving car in the next five years and 43% feel driverless cars are safer.

based scene understanding models to automate the analysis process. Besides road infrastructure, communication infrastructure must be extended to support new communication technologies such as 4G and 5G networks as well as V2I and V2V systems.

In the Arab region, UAE was recently ranked within the world top 10 countries in automated vehicles readiness.⁸ Moreover, Egypt is currently installing a unified network of cameras on roads that monitors traffic nationwide to help regulate drivers' behavior and improve road infrastructure. According to the World Economic Forum's 2019 road-quality report, UAE and Oman ranked 8th and 10th. Furthermore, the progress is noticeable—in a five-year period, Egypt's ranking jumped from 29 to 118, and Algeria moved from 107 to 67.

Much of the problem of AD development is the need for a significant amount of training data to train complex machine learning algorithms, while some events are rare like witnessing an accident. Unfortunately, the overwhelming majority of the data collection efforts worldwide are not recorded in cities with chaotic driving. The challenging driving scenes created by the aforementioned road infrastructure challenges, in addition to unconventional traffic composition as captured in Figure 2, are not represented in the overwhelming majority of existing datasets. The problem can be illustrated briefly by imagining how many times such models will cause the vehicle to stop due to pedestrian false positive detections caused by passengers in the back of tricycles or pickup trucks, as these situations were not

provided during the training time. In parts of the Arab region today, equipping a laser scanner (LiDAR) on a car is not allowed on public roads for national security reasons.

Applying transfer learning from the available models trained already on abundance of data from the developed countries' roads can be a solution to the data scarcity challenge. Key solutions for the need for 'long-tailed' distributions of data and handling imbalanced data distributions are to rely on synthetic data generation, multi-task learning, and to apply class-weighting methods and develop realistic simulators capturing the nature of driving in the region. Companies and universities data collection initiatives should be encouraged.

Legislation and culture challenges. In the autonomous era, the driving task is shared between the driver and the vehicle, yet accident liability is currently assigned to the driver. With the introduction of AD, regulations in developed countries started to adapt to the capability to self-drive without being actively controlled nor monitored by a human.

Accident liability started to consider automakers, and it is difficult to expect how automakers will respond to the regulations in developing countries with higher accident rates and traffic rules violations. Google's Waymo autonomous cars, the world's leader in miles driven, have driven 1/5th of



Figure 1. Common unofficial pedestrian waiting areas on highways in much of the Arab region.

the average human driving miles before a fatal accident.⁸ The number is a clear indicator that AD testing is far from proving it is safer than human driving, the regulatory safety standards in the Arab region, except Dubai, are not defined and, unfortunately, the developed regulations worldwide are neither unified nor directly applicable elsewhere.

Bureaucracy and leniency in law enforcement in parts of the Arab region could make it more complex to reach a satisfactory legislation for all stakeholders. Regulations should aim to precautionarily reduce the chances of creating dangerous driving situations. One example from the Arab region highlighting such necessity is the lack of standards governing highway billboards specifications, as in Figure 3 sample, which represents a serious source of driver distraction. Other examples include the

leniency in following heavy vehicles allowed driving times regulations, driving in the opposite directions, unauthorized parking, and the lack of pedestrians' awareness about traffic regulations.

Accordingly, the Arab countries must start conversations among regulators, the public, and stakeholders about how autonomous vehicles and its operation should be regulated. It is important to build on top of the lessons learned from the developed countries legislation initiatives, such as the U.K.'s Automated and Electric Vehicles Act, and to aim for a unified regulatory framework across the region from the beginning. It is helpful to target a degree of consistency from the start given that vehicles are potentially capable of moving across borders. UAE started legislation of self-driving cars testing and accident responsibility. Such initia-



Figure 2. Unconventional traffic composition from the Arab region highways; a passenger tricycle, an auto rickshaw, an animal-driven vehicle, and a passenger pickup truck.



Figure 3. Examples depicting the lack of highway regulations, including billboard overload, missing lane markings, and no laws governing driving behind large trucks carrying heavy cargo.

tive can be an important guidance to the neighboring countries in the region to start. The defined regulatory safety standards should be flexible in terms of the data requirements for safety authorization. Moreover, the legislative approach should be clear and fair regarding accident liability and ‘recklessness’ identification. Bureaucracy should be reduced and testing of self-driving cars should be encouraged. Recently, Brightskies was authorized to test autonomous cars in Egypt and managed to reveal their first AD system. Volkswagen and the Qatar Investment Authority signed Project Qatar Mobility that aims to create a comprehensive ecosystem for AD, including the legal framework and infrastructure. Similar Initiatives like that should be encouraged across the region.

Innovation and technology challenges. A strong innovation ecosystem enables attracting investment and building the infrastructure for adopting AD. Despite the critical need for innovation in developing countries and the majority of the Arab region, they invest far less compared to advanced coun-

tries.² The governmental sector, the private sector, and even the civil society must cooperate to encourage launching more initiatives for propelling innovation and to ensure sustainable investments in R&D. This includes educational programs at universities, participation in international scientific and industrial events, and promoting collaborative research efforts.

A promising example from the region is UAE, it ranked 34th in the 2020 Global Innovation Index. However, rankings for other Arab countries indicate they still have serious challenges. Regarding industry and academia collaboration, a successful example is conducted by KAUST University that launched self-driving shuttles on its campus to enable students and researchers to collaborate with the industry. Dubai’s Roads and Transport Authority launched the second edition of the World Self-Driving Transport Challenge 2021, where firms and universities collaborate. Additionally, Valeo as a leader in the automotive industry, maintains several programs with universities in Egypt to

bridge the gap between academia and industry through conducting funded joint research projects and developing technical courses. Valeo Egypt contributed to creating the healthy echo system that initiated automotive embedded software services start-ups in Egypt like Av-elabs, eJad, AvidBeam, and Brightskies.

Customer acceptance challenges. The perception of autonomous vehicles varies in different regions based on culture, the level of technology awareness, and social interactions.¹ Tailored awareness campaigns should be started targeting the public in Arab countries considering their needs and income levels. Governments and auto-

makers should cooperate to encourage the adoption of autonomous vehicles through joint financial support programs. An example of effective introduction of the technology from the Arab region is the initiative of launching a fleet of self-driving electric shuttles in the 2022 FIFA World Cup in Qatar (see Abbar et al. on p. 67). According to a recent survey, approximately half of UAE residents are likely to own a self-driving car in the next five years and 43% feel driverless cars are safer. The region market is already attracting key stakeholders, as example, Neolix signed an agreement to test autonomous vehicles in Saudi Arabia and UAE customized for the region’s weather conditions. AutoX aims to launch the region’s first commercial ‘robotaxi’ and autonomous delivery service in UAE. Those examples show that despite the challenges, there are very promising opportunities for adopting self-driving technologies in the Arab region.

Conclusion

Unconventional challenges are facing the adoption of AD in the Arab region related to infrastructure, legislation, technology, and consumer acceptance. Those challenges threaten to deprive the region of the great benefits of such technology despite the desperate need and the great commercial value for carmakers. As a step to accelerate AD adoption in the Arab region, this article sheds light on those challenges and provided preliminary solutions building on existing efforts in the region. AD deployment should be considered a strategic goal that requires collaborative efforts from all stakeholders. 

References

1. Cho, E. and Jung, Y. Consumers’ understanding of autonomous driving. *Information Technology & People* (2018).
2. Cirera, X. and Maloney, W.F. The innovation paradox: Developing-country capabilities and the unrealized promise of technological catch-up. The World Bank, 2017.
3. Eraqi, H.M., Moustafa, M.N., and Honer, J. Conditional imitation learning driving considering camera-LiDAR fusion and efficient occupancy grid mapping. In *Proceedings of the IEEE 23rd Intern. Conf. Intelligent Transportation Systems* (Greece, Sept. 20–23, 2020).
4. Hassan, A.M., Abdel-Nasser, K.G., and Saied, A.M. Ahmed, A. Modeling and valuation of traffic accidents on rural roads in Egypt. Master’s thesis, Cairo University, 2004.
5. International Energy Agency. Global car sales by key markets, 2005–2020; <https://www.globalinnovationindex.org/gii-2020-report>.
6. Janai, J., Guney, F., Beht, A. and Geiger, A. Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art, 2017; arXiv:1704.05519.
7. Threlfall, R. Autonomous vehicles readiness index. KPMG International, 2020.
8. Waymo. Safety Report, 2020; <https://waymo.com/safety/>.
9. World Health Organization. Global status report on road safety 2018; <https://apps.who.int/iris/bitstream/handle/10665/276462/9789241565684-eng.pdf>.

Hesham M. Eraqi is Senior Expert of AI and Deep Learning at Valeo and Adjunct Faculty in the Computer Science and Engineering Department at the American University in Cairo, Egypt.

Ibrahim Sobh is Senior Expert of AI and Deep Learning at Valeo, Cairo, Egypt.

Traffic Routing in the Ever-Changing City of Doha

BY SOFIANE ABBAR, RADE STANOJEVIC, SHADAB MUSTAFA, AND MOHAMED MOKBEL

ON DECEMBER 2, 2010, Qatar was announced to host 2022 FIFA World Cup.

That was time for celebrating the first-ever Middle Eastern country to organize the tournament. The 1.8M population of Qatar then (2.8M today) never imagined the journey their country was about to embarked. Indeed, in less than 10 years, the population grew by more than a half, pushing the available urban resources and services to their limit. At the same time, the country undertook an ambitious investment plan of \$200B on various infrastructural projects including a brand new three-line metro network, six new stadiums, several new satellite cities, and an astonishing 4,300km of new roads, which tripled the size of the road network in only five years.³

While this enterprise boosted the socio-economical life of people in Qatar, it did disrupt the way they navigate the urban space and their mobility patterns in general. Simple commutes to work, drops and pickups of kids to and from schools, became challenging and impossible to plan with daily changes in the road layout, including temporary and permanent closures, deviations, new connections, conversions of roundabouts into signaled intersections, turn restrictions, to name but a few. A commute to school

that lasted 10 minutes yesterday, could last 25 minutes today. Cab drivers in the city of Doha (Qatar's capital), who are mostly foreigners, also wish they could rely on popular navigation services such as Google Maps, Here, or Tomtom.

Yet, all such systems fall short in coping up with the rapid urbanization and the ever-changing roads in Doha. This was actually depicted in a very popular caricature in one of the most widely distributed daily local newspapers showing Google maps as a limping turtle that is helplessly trying to catch a bunny representing the road changes in the city of Doha.⁴

Besides the general public who is not happy with the routes offered by navigation systems, other stakeholders from public and private sectors were struggling with the poor quality of existing digital maps. For example, the Ministry of Transport and Communication was facing issues getting access to the most accurate map of the road network, needed for their traffic modeling. Also, transportation, delivery, and logistics companies that heavily rely on accurate maps, routes, and travel time estimates were tired of the many lost drivers and missed rendezvous.

Early work: Silent maps are not enough. The issue of inaccurate local maps has triggered an early work at Qatar Computing Research Institute (QCRI) in collabora-



tion with Qatar Mobility Innovation Center (QMIC) to come up with an accurate map for the city of Doha, Qatar.⁷ The idea was to use data collected from a fleet of vehicles that are continuously tracked, for accurate and timely detection of road changes, such as new roads, road closures, and detours. Though that early work was successful in coming up with a more accurate map than what navigation systems have, it was not enough to address the main problem of routing. Accurate topological maps do not say much about the time needed to go through each road segment—a main functionality needed for any routing application.

Data access and collaboration. To address the routing problem in the ever-changing roads of Doha, we partnered with the national taxi company Karwa. The collaboration gives us access to all taxi data (both historic and live) that took place in the country, including pick-up and drop-off locations, time, duration,

speed, fare, route, as well as sampled GPS points for each trip—a gold mine for our research agenda. But most importantly, we also learned from our partners about the real challenges they face, which helped us prioritize our projects.

Map enrichments for traffic-aware routing. Our first project with Karwa was to enrich the topological maps with traffic information, that is, accurate edge weights for each road segment for each hour of the day. Inferring traffic information from a large number of vehicles can be relatively straightforward. However, the problem is much more challenging when the data is sparse and does not cover many roads with large frequency. We tackle these problems in Stanojevic et al.^{5,6} and derive a traffic layer with an accuracy comparable to the commercial maps using only sparse data available to us either from Karwa Taxi data as in Stanojevic et al.⁵ or from using commercial map APIs as in Stanojevic et

Our research led to the deployment of a system in production for two large companies making over six million routing and travel time requests per month.

al.⁶ In particular, we devise a supervised learning approach for inferring the travel times on the individual road segments using a cohort of trajectories with known travel times. To tackle the issue of data sparsity, we group road segments that are likely to have similar road speeds at a given time. In order to ensure our inferred traffic model accurately models the actual road conditions, we explicitly constrain the solution to the problem using the available metadata and Ridge regularization. For example, a road segment cannot have an average speed that is negative or larger than its speed limit. The resulting machine learning framework allows us to accurately detect the traffic model on every road segment in the city (including many that have little or no data). We validate our model using not only out-of-sample testing, but also by continuously evaluating the accuracy of the model in the deployed system over hundreds of thousands of trajectories every week.

Indeed, our proposal led to the creation of a new traffic-aware routing system, named QARTA. To evaluate the quality of our system, we designed a comparison of QARTA API vs. Google Maps API, in which we requested travel-time predictions from both systems for ~200K real taxi trips before they took

place. By using an up-to-date map of the city and accurate traffic models, our system achieved a 20% reduction in median absolute errors compared to Google Maps API. This gain is quite substantial at scale. Indeed, for a taxi or delivery company that makes 1M trips/month with an average trip duration of 15 minutes, saving one minute on each trip can lead to virtually 6.66% operational efficiency improvement either by increasing the number of trips or by reducing the waiting time.

Travel time estimation.

Another interesting line of work is concerned with accurate travel-time predictions which is at the core of the taxi and delivery industry. There are several instances where companies need travel time without routes. For example, the case of driver dispatching where a driver is selected among others based on proximity to the target. Thus, we came up with STAD, which stands for Spatio-Temporal ADjustment of travel time.¹ The idea behind STAD is as follow: Let say Layla wants to go from the Pearl neighborhood to Lusail stadium for a soccer game on Thursday night. A free-flow routing engine can spit a route and a travel time of 16 minutes, based on roads free-flow speed. However, if we have access to historical trips in the city, with

their origin and destination locations, departure time, and duration, then it should be possible for us to learn an adjustment factor of the free-flow travel time, by taking into account the spatial localization of the origin and destination points, as well as the departure time and day, leading to a more accurate travel time estimation of 23 minutes. In our implementation, we partition the city into small area zones (that is, block polygons) and map each location (origin or destination) to its corresponding block that we use as spatial features. We also partition the time into hours of the day and days of the week and we use as temporal features. Finally, we devise a supervised regression model based on gradient boosting machines to learn the adjustment factor for different spatio-temporal combinations of trips. Our experiments on 750K trips in Doha show that STAD API achieves a median absolute error of 126 seconds against 146 seconds for Google Maps, yielding an error reduction of 14%.

From research to practice.

While tackling the research questions discussed here was academically exciting, making true impact required deploying our solution and putting it in the end-user's hands. To that end, we built QARTA, a traffic-aware digital map engine, supporting enterprise customers, for example, transportation and logistics companies, via routing APIs. Currently, we have two major customers of QARTA API that use our technology in production. The first is Karwa taxi company for which we resolve around 1M routing and travel time requests per week; the second is Rafeeq delivery company which makes over 750,000 requests

per week to our system, yielding improved QoS and significant financial savings compared to the commercial solution they previously used. We are currently actively exploring partnerships with the adjacent logistics sector (food and last-mile delivery) where accurate route and travel-time predictions are of paramount value for efficient operation.

Next steps. Qatar is not the only fast-growing country, as many countries and cities in the Middle East, Africa, and Southern Asia are also significantly expanding their infrastructure,² creating a real opportunity for more functional and fast-updating digital maps. We plan to take our system QARTA beyond local impact in Qatar to other international markets. 

References

1. Abbar, S., Stanojevic, R. and Mokbel, M. STAD: Spatio-temporal adjustment of traffic-oblivious travel-time estimation. In *Proceedings of the IEEE Intern. Conf. Mobile Data Mgmt.* (Versailles, France, June 2020).
2. Abyad, A. Demographic changes in the GCC countries: Reflection and future projection. *Middle East J. Age and Ageing*, 2018.
3. *The Peninsula*. Qatar road network increased three times between 2013-18: Ashghal; <http://bit.ly/2KArihY>
4. *Raya Daily* (Sept. 8, 2020), 20; <https://bit.ly/3pryzk>
5. Stanojevic, R., Abbar, S. and Mokbel, M. W-edge: Weighing the edges of the road Network. In *Proceedings of the ACM SIGSPATIAL Intern. Conf. Advances in GIS*, (Seattle, WA, USA, Nov. 6–9, 2018).
6. Stanojevic, R., Abbar, S. and Mokbel, M. MapReuse: Recycling routing API queries. In *Proceedings of the IEEE Intern. Conf. Mobile Data Mgmt.* (Hong Kong, China, June 2019).
7. Stanojevic, R., Abbar, S., Thirumuruganathan, S., Chawla, S., Filali, F., and Aleimat, A. robust road map inference through network alignment of trajectories. In *Proceedings of the SIAM Intern. Conf. on Data Mining*, (San Diego, CA, USA, May 2018).

Sofiane Abbar, Qatar Computing Research Institute, HBKU, Doha, Qatar.

Rade Stanojevic, Qatar Computing Research Institute, HBKU, Doha, Qatar.

Shadab Mustafa, Karwa Technologies, Mowasalat, Doha, Qatar.

Mohamed Mokbel, Qatar Computing Research Institute, HBKU, Doha, Qatar.

Copyright held by authors/owners. Publication rights licensed to ACM.

ArabHCI: Five Years and Counting

BY SHAIMAA LAZEM, MENNATALLAH SALEH, AND EBTISAM ALABDULQADER

ARABHCI IS AN initiative that started in 2016 to promote Human-Computer Interaction (HCI) research and education in Arab countries, and to build a community of Arab and non-Arab researchers interested in the Arab context (<https://arabhci.org>). Notably, the inception team consisted of all Arab female researchers.

The Arab world consists of 22 countries across Asia and Africa and is considered one of the world's most strategic territories to host renowned political events such as the Arab Spring and the refugee crisis; all of which featured unique appropriations of existing social media technologies. The initiation of ArabHCI was largely motivated by concerns about how Arab users have been sometimes misrepresented in global HCI research that focused on



Attendees of the CHI'19 Diversity Lunch held in Glasgow, Scotland.

these events. The diversity and cultural richness of the region were not fully communicated to Western technology makers. Therefore, our research agenda was set to increase the visibility of local HCI researchers' perspectives and expertise and to explore the methodological means by which the authentic voices of Arab users could be included in the technology design processes.²

HCI courses are gaining more popularity in Arab academic institutions despite the fact that HCI's interdisciplinary nature and its link to humanities and art make it challeng-

ing for computer science educators to introduce it to students. One of ArabHCI goals is to promote a broader and deeper teaching agenda that goes beyond a focus on user interface guidelines. HCI concepts, we believe, could be integrated when students are taught about emerging technologies such as wearable devices, virtual/augmented reality, brain-computer interfaces, and autonomous vehicles. A smooth interaction design is essential to the success of these technologies. Additionally, to help students grasp the different means by which users could be engaged in the design, it is important to introduce design methods as well as quantitative and qualitative research to HCI curricula.

Since its inception, ArabHCI has embarked on a journey to establish an intercultural dialogue between the global HCI community—a.k.a. Western technology designers—and Arab HCI

researchers. At the core of this dialogue is our commitment to communicate the contextual challenges hindering Arab researchers from participating in the global HCI community to the ACM Special Interest Group on Computer-Human Interaction (SIGCHI). We contributed to SIGCHI's decisions of providing a reduced fee for emerging economies and offering an early career researcher mentorship program. ArabHCI's inauguration and success as a regional initiative were featured twice in the diversity and inclusion events at the SIGCHI flagship conference in 2017 and 2019.

Furthermore, four research-focused meetings were hosted at top HCI international conferences that encouraged an exchange of views between Arab and non-Arab HCI researchers. The 46 papers presented in those workshops depicted HCI research in 10 Arab countries: Saudi Arabia,

Cultural values must be incorporated within systems to ensure their adaptability and success.



About

ArabHCI, an initiative established in September 2016, aims to empower, bridge, and connect HCI researchers and practitioners from the Arab world with those who are conducting/interested in research in this context and ACM SIGCHI. The community goal is to leverage our "insider" understanding of HCI research in the Arab context and explore the challenges and unique opportunities for future research.



Members

ArabHCI community is open for all HCI researchers interested in the Arab context with no restrictions on their background, gender, or nationality.

Our Events



Half-Day Workshop



Envision a research and practice agenda for the community



Arab and Non-Arab HCI Scholars

Themes of Discussion

Theme 1: Current Research Status in the Arab World

There are various challenges that can hinder the work researchers do in the Arab region. These include safety, security and political challenges as well as movement and ownership concerns.

Future recommendations include establishing meaningful partnerships with governments, NGOs and local initiatives.

Theme 2: Methodologies for Research with Arab Cultures

The unique context of the Arab world can create interesting challenges for researchers conducting qualitative fieldwork. Despite these challenges, there is a lack of methodological reporting and reflection as to how to conduct studies in the region.

Future recommendations include encouraging HCI researchers and practitioners to critically reflect and share their experiences via (e.g. conferences, white papers and magazines).

Theme 3: Appropriating versus Designing New Technologies for Arab Cultures

There are many factors that researchers consider when attempting to achieve a culturally sensitive-design. However, there are many challenges that stand in the way of researchers to best determine and make use of these various elements.

Future recommendations include taking an inclusive approach and creating a creative appropriation agenda.

Egypt, Iraq, Kuwait, Palestine, Lebanon, Morocco, Jordan, Syria, and Qatar. Non-Arab researchers participated from the U.S., U.K., Germany, Denmark, Australia, Austria, Belgium, Canada, and Iran. Following the success of these workshops, ArabHCI members were invited to edit a special topic to promote their work with Arab users to the global audience in ACM *Interactions* magazine.⁵

Distinctive ArabHCI Research

Community discussions revealed some underlying Western assumptions about Arab users that do not translate to reality. Thus, it is important to establish the unique characteristics of the Arab communities that distinguish Arab from Western users and design digital technologies to cater to them. Here, we discuss efforts by Arab and non-Arab researchers to highlight some of the appropriations of Western design methods they had to make to suit the Arab context.

Assumptions and realities. International researchers, when adapting their systems to Arabic, have often simply translated the interfaces. While this is usable, it does not provide for a quality user experience. Cultural values must be incorporated within systems to ensure their adaptability and success. For example, when designing social networking websites, the English word "friend" has 26 different translations in Arabic, choosing which translation to use in context influences the design acceptance.

Most Arabs are very in

touch with their religion and universal designs do not accommodate for that. Some Arab users reported they do not mark a Facebook event as "going" and instead use "interested" even if they are certain they will attend. The provided explanation was that all future events in their offline life are followed by the sentence "God willing," which is an option that is not reflected online through Facebook. Religion is a significant driver for Arab Muslims' behavior as it appeared in how they integrate it into their online practices; for example, using the Holy Quran during tweets.¹ This behavior, after investigation, was interpreted as an act of worship since social media accounts could extend beyond their lifetime and so are considered good places to share good deeds.

Another significant marker of Arab cultures is gender differences, while this might be known, the details of how it influences technology use is not incorporated in Western designs. Saleh et al.⁹ reported that Arab women are concerned with their reputation and have greater requirements for privacy protection of their personal data. All these considerations and more only touch the surface of the depth and richness of the Arab culture. It is also important to note that not all Arab countries are the same in terms of cultural practices, there is a large spectrum from conservative cultures like Saudi Arabia to secular cultures like Lebanon.² Designers should be cautious about generalizing lessons learned from one country to another.

A poster depicting the vision of and events hosted by ArabHCI.

Unique design methods for Arab users. While technology design has established Western-originated methods, researchers encountered several challenges when implementing them in the Arab context. Giglitto et al.³ found that when using participatory design methods with Egyptian Bedouins, the participants were constantly cautious about making mistakes that will portray them as incompetent, as opposed to Western participants who may enjoy the process of prototyping. What appeared as extra cautiousness stems from a culture that is grounded in problem avoidance since a capable Bedouin would secure grazing areas in the desert before bringing the sheep. These findings were echoed by Saleh and Sturm¹⁰ during their work with low-literate Egyptians on a simple oral survey. Their participants were hesitant to answer questions in fear of giving wrong answers despite being assured there were no right or wrong answers.

Another challenge faced by Nassir and Leong⁸ was the difficulty encountered during conducting in-home interviews due to the busy nature of the Saudi homes and the large number of people in the home. To overcome these challenges, the authors introduced cultural probes that have assisted them with this unique Arab problem. Other methods were employed such as diaries, social media as a communication channel between participants and researchers, and seeking the help of chaperons as co-researchers.

Another unique take to a rather prominent problem is participant recruitment. While normally this is challenging for any researcher, participant recruitment in the Arab world has added complexity, especially when recruiting female participants. Nassir et al.⁷ discuss this extensively, noting that in the conservative Saudi society, the male guardians of female participants may not approve of their communication even with female researchers. This issue has been reported across multiple studies and is usually quite challenging to resolve.

Last but not least, the Arab context has unexplored design spaces to which technology must be tailored. One example is the annual Hajj pilgrimage where millions of Muslims visit Makkah for a few weeks. This distinctive gathering has special arrangements and therefore special requirements should be investigated on the ground as shown by Majrashi,⁶ who proposed unique technology solutions. Another impactful issue is the work done with refugees to explore their use of digital technologies. One of the noteworthy examples of this work is conducted by members of the ArabHCI community to investigate participatory design challenges with such marginalized communities.⁴

In the presence of such complex socio-technical system dynamics, an insider perspective can be very helpful in adapting existing research methodologies and generating results with and for the Arab communities.

The Arab context has unexplored design spaces to which technology must be tailored.

Conclusion

Thanks to our small community, Arab students and faculty interested in learning about HCI have a platform to connect with established HCI researchers. ArabHCI community's collective knowledge is a useful resource for Western designers who wish to scrutinize their assumptions and methods before conducting research in the region. It could be leveraged to increase their familiarity with the context if they were to judge work done by Arab scholars. We plan to add more high-quality articles to our resources through editing special issues in top HCI journals. Additionally, future efforts will focus on organizing local events to grow technical HCI capacity in Arab countries.

Acknowledgments. We thank all ArabHCI workshop organizers, workshop participants, and the global HCI community members for actively engaging with us. 

References

1. Abokhodair, N., Elmadany, A., and Magdy, W. Holy tweets: Exploring the sharing of the Quran on Twitter. In *Proceedings of the ACM Hum.-Comput. Interact.* 4, Article 159 (Oct. 2020); <https://doi.org/10.1145/3415230>
2. Alabdulqader, E., Abokhodair, N., and Lazem, S. Human-computer interaction across the Arab world. In *Proceedings of the 2017 CHI Conf. Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1356–1359; <https://doi.org/10.1145/3027063.3049280>
3. Giglitto, D., Lazem, S., and Presto, A. In the eye of the student: An intangible cultural heritage experience, with a human-computer interaction twist. In *Proceedings of the 2018 CHI Conf. Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 290, 1–12; <https://doi.org/10.1145/3173574.3173864>
4. Krüger, M., Duarte, A.B., Weibert, A., Aal, K., Talhouk, R., and Metatla, O. What is participation? Emerging challenges for participatory design in globalized conditions. *Interactions* 26, 3 (May–June 2019), 50–54; <https://doi.org/10.1145/3319376>
5. Lazem, S., Alabdulqader, E., and Khamis, M. Introduction to ArabHCI Special Topic. *Interactions* 26, 3 (May–June 2019), 41–43; <https://doi.org/10.1145/3320109>
6. Majrashi, K. User need and experience of Hajj mobile and ubiquitous systems: Designing for the largest religious annual gathering. *Cogent Engineering* 5, 1 (2018); <http://doi.org/10.1080/23311916.2018.1480303>
7. Nassir, S., Al-Dawood, A., Alghamdi, E., and Alyami, E. 'My guardian did not approve!' Stories from fieldwork in Saudi Arabia. *Interactions* 26, 3 (May–June 2019), 44–49; <https://doi.org/10.1145/3318145>
8. Nassir, S. and Leong, T.W. Traversing boundaries: Understanding the experiences of ageing Saudis. In *Proceedings of the 2017 CHI Conf. Human Factors in Computing Systems*. ACM, New York, NY, USA, 6386–6397; <https://doi.org/10.1145/3025453.3025618>
9. Saleh, M., Khamis, M., and Sturm, C. What about my privacy, Habibi? INTERACT 2019. D. Lamas, F. Loizide, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris (eds). LNCS 11748. Springer, Cham; https://doi.org/10.1007/978-3-030-29387-1_5
10. Saleh, M., and Sturm, C. Exploring the effect of literacy on signs in GUI design. In *Proceedings of the 2nd African Conf. Human-Computer Interaction: Thriving Communities* (2018). ACM, New York, NY, USA, Article 21, 1–5; <https://doi.org/10.1145/3283458.3283534>

Shaimaa Lazem is an associate professor for the City of Scientific Research and Technology Applications (SRTA-city) Alexandria, Egypt.

Mennatallah Saleh is a Steering Committee member of the Cairo ACM SIGCHI Chapter (CairoCHI), Egypt.

Ebtisam Alabdulqader is an assistant professor for CCIS, King Saud University Riyadh, Saudi Arabia.

BY KAREEM DARWISH, NIZAR HABASH, MOURAD ABBAS, HEND AL-KHALIFA, HUSEEIN T. AL-NATSHEH, HOUDA BOUAMOR, KARIM BOUZOUBAA, VIOLETTA CAVALLI-SFORZA, SAMHAA R. EL-BELTAGY, WASSIM EL-HAJJ, MUSTAFA JARRAR, AND HAMDY MUBARAK

A Panoramic Survey of Natural Language Processing in the Arab World

THE TERM *NATURAL language* refers to any system of symbolic communication (spoken, signed, or written) that has evolved naturally in humans without intentional human planning and design. This distinguishes natural languages such as Arabic and Japanese from artificially constructed languages such as Esperanto or Python. Natural language processing (NLP), also called computational linguistics or human

language technologies, is the sub-field of artificial intelligence (AI) focused on modeling natural languages to build applications such as speech recognition and synthesis, machine translation, optical character recognition (OCR), sentiment analysis (SA), question answering, and dialogue systems. NLP is a highly interdisciplinary field with connections to computer science, linguistics, cognitive science, psychology, mathematics, and others.

Some of the earliest AI applications were in NLP (machine translation, for example); and the last decade (2010–2020) in particular has witnessed an incredible increase in quality, matched with a rise in public awareness, use, and expectations of what may have seemed like science fiction in the past. NLP researchers pride themselves on developing language-independent models and tools that can be applied to all human languages. Machine translation systems, for example, can be built for a variety of languages using the same basic mechanisms and models. However, the reality is that some languages (English and Chinese) do get more attention than others (Hindi and Swahili). Arabic, the primary language of the Arab world and the religious language of millions of non-Arab Muslims, is somewhere in the middle of this continuum. Though Arabic NLP has many challenges, it has seen many successes and developments.

Next, we discuss Arabic's main challenges as a necessary background, and we present a brief history of Arabic NLP. We then survey a number of its research areas, and close with a critical discussion of the future of Arabic NLP. An extended version of this article including almost 200 citations and links is on Arxiv.^a

Arabic and Its Challenges

Arabic today poses a number of modeling challenges for NLP: morphological richness, orthographic ambiguity,

^a <https://arxiv.org/abs/2011.12631>





dialectal variations, orthographic noise, and resource poverty. We do not include issues of right-to-left Arabic typography, which is an effectively solved problem (although not universally implemented).

Morphological richness. Arabic words have numerous forms resulting from a rich inflectional system that includes features for gender, number, person, aspect, mood, case, and a number of attachable clitics. As a result, it is not uncommon to find single Arabic words that translate into five-word English sentences: *wa+sa+ya-drus-uuna+ha* 'and they will study it.' This challenge leads to a higher number of unique vocabulary types compared to English, which is challenging for machine learning models.

Orthographic ambiguity. The Arabic script uses optional diacritical marks to represent short vowels and other phonological information that is important to distinguish words from each other. These marks are almost never used outside of religious texts and children's literature, which leads to a high degree of ambiguity. Educated Arabs do not usually have a problem with reading undiacritized Arabic, but it is a challenge for Arabic learners and computers. This out-of-context ambiguity in Standard Arabic leads to a staggering 12 analyses per word on average: for example, the readings of the word *كتبتك* *ktbt* (no diacritics) includes *كاتبتك* *katabtu* 'I

wrote,' *كاتبتك* *katabat* 'she wrote,' and the quite semantically distant *كاتبتك* *ka+t~ibit* 'such as Tibet.'

Dialectal variation. Arabic is also not a single language but rather a family of historically linked varieties, among which Standard Arabic is the official language of governance, education, and the media, while the other varieties, so-called dialects, are the languages of daily use in spoken, and increasingly written, form. Arab children grow up learning their native dialects, such as Egyptian, Levantine, Gulf, or Moroccan Arabic, which have their own grammars and lexicons that differ from each other and from Standard Arabic. For example, the word for 'car' is *قرايس* *syArp* (*sayyaara*) in Standard Arabic, *قرايس* *Erbyp* (*arabiyya*) in Egyptian Arabic, *قربك* *krhbp* (*karhba*) in Tunisian Arabic, and *رموتوم* *mwtr* (*motar*) in Gulf Arabic. The differences can be significant to the point that using Standard Arabic tools on dialectal Arabic leads to quite sub-optimal performance.

Orthographic inconsistency. Standard and dialectal Arabic are both written with a high degree of spelling inconsistency, especially on social media: A third of all words in Modern Standard Arabic (MSA) comments online have spelling errors; and dialectal Arabic has no official spelling standards, although there are efforts to develop such standards computationally, such as the work on CODA, or Conventional Orthography for

Dialectal Arabic. Furthermore, Arabic can be encountered online written in other scripts, most notably, a Romanized version called Arabizi that attempts to capture the phonology of the words.

Resource poverty. Data is the bottleneck of NLP; this is true for rule-based approaches that need lexicons and carefully created rules, and for machine learning (ML) approaches that need corpora and annotated corpora. Although Arabic unannotated text corpora are quite plentiful, Arabic morphological analyzers and lexicons as well as annotated and parallel data in non-news genre and in dialects are limited.

None of the issues mentioned here are unique to Arabic—for example, Turkish and Finnish are morphologically rich; Hebrew is orthographically ambiguous; and many languages have dialectal variants. However, the combination and degree of these phenomena in Arabic creates a particularly challenging situation for NLP research and development. Additional information has been published on Arabic computational processing challenges.^{4,5}

A Brief History of NLP in the Arab World

Historically, Arabic NLP can be said to have gone through three waves. The first wave was in the early 1980s with the introduction of Microsoft MS-DOS 3.3 with Arabic language support. In 1985, the first Arabic morphological analyzer was developed by Sakhr Software. Most of the research in that period focused on morphological analysis of Arabic text by using rule-based approaches. Sakhr has also continued leading research and development in Arabic computational linguistics by developing the first syntactic and semantic analyzer in 1992 and Arabic optical character recognition in 1995. Sakhr also produced many commercial products and solutions, including Arabic-to-English machine translation, Arabic text-to-speech, and an Arabic search engine. This period almost exclusively focused on Standard Arabic with a few exceptions related to work on speech recognition.

The second wave was during the

years 2000-2010. Arabic NLP gained increasing importance in the Western world especially after September 11. The U.S. funded large projects for companies and research centers to develop NLP tools for Arabic and its dialects, including machine translation, speech synthesis and recognition, information retrieval and extraction, text-to-speech, and named entity recognition. Most of the systems developed in that period used machine learning, which was on the rise in the field of NLP as a whole. In principle, ML required far less linguistic knowledge than rule-based approaches and was fast and more accurate. However, it needed a lot of data, some of which was not easy to collect, for example, dialectal Arabic to English parallel texts. Arabic's rich morphology exacerbated the data dependence further. So, this period saw some successful instances of hybrid systems that combine rule-based morphological analyzers with ML disambiguation which relied on the then newly created Penn Arabic Treebank (PATB). The leading universities, companies, and consortia at the time were Columbia University, the University of Maryland, IBM, BBN, SRI, the Linguistic Data Consortium (LDC), and the European Language Resources Association (ELRA).

The third wave started in 2010, when the research focused on Arabic NLP came back to the Arab world. This period witnessed a proliferation of Arab researchers and graduate students interested in Arabic NLP and an increase in publications in top conferences from the Arab world. Active universities include New York University Abu Dhabi (NYUAD),^b American University in Beirut (AUB), Carnegie Mellon University in Qatar (CMUQ), King Saud University (KSU), Birzeit University (BZU), Cairo University, and others. Active research centers include Qatar Computing Research Institute (QCRI),^c King Abdulaziz City for Science and Technology (KACST), and more. It should be noted that there are many actively contributing researchers in smaller groups across the Arab world. This pe-

riod also overlapped with two major independent developments: the rise of deep learning and neural models, and the rise of social media. The first development affected the direction of research, pushing it further into the ML space; the second led to the increase in social media data, which introduced many new challenges at a larger scale: more dialects and more noise. This period also witnessed a welcome increase in Arabic language resources and processing tools, and a heightened awareness of the importance of AI for the future of the region—for example, the UAE now has a ministry specifically for AI. Finally, new young and ambitious companies such as Mawdoo3 are competing for a growing market and expectations in the Arab world.


Arabic Tools and Resources

We organize this section on Arabic tools and resources into two parts: first, we discuss enabling technologies which are the basic resources and utilities that are not user-facing products; and second, we discuss a number of advanced user-targeting applications.


Resource construction is a lengthy and costly task that requires significant teamwork among linguists, lexicographers, and publishers over an extended period of time.

Corpora. NLP relies heavily on the existence of corpora for developing and evaluating its models, and the performance of NLP applications directly depends on the quality of these corpora. Textual corpora are classified at a high level as annotated and unannotated corpora.

Annotated corpora are a subset of unannotated corpora that have been enriched with additional information such as contextual morphological analyses, lemmas, diacritizations, part-of-speech tags, syntactic analyses, dialect IDs, named entities, sentiment, and even parallel translations. The more information, the costlier the process is to create such corpora. For Arabic, the main collections of annotated corpora were created in its second wave, mostly outside the Arab world. The most notable annotated resource is the LDC's Penn Arabic Treebank (PATB), which provides




The success of word embedding models trained on unannotated data and resulting in improved performance for NLP tasks with little or no feature engineering has led to many contributions in Arabic NLP.




b <http://www.camel-lab.com>

c <https://alt.qcri.org/>



Among the challenges facing Arabic NER is the lack of letter casing, which strongly helps English NER, and the high degree of ambiguity, including especially confusable proper names and adjectives.



a relatively large MSA corpus that is morphologically analyzed, segmented, lemmatized, tagged with fine-grained parts of speech, diacritized, and parsed. PATB has enabled much of the Arabic NLP research since its creation. The Prague Arabic Dependency Treebank (PADT) was the first dependency representation treebank for Arabic. The Columbia Arabic Treebank (CATiB) was an effort to develop a simplified dependency representation with a faster annotation scheme for MSA. The University of Leeds' Quranic Arabic Corpus is a beautifully constructed treebank that uses traditional morpho-syntactic analyses of the Holy Quran. With the rising interest in dialectal data, there have been many efforts to collect and annotate dialectal data. The LDC was first to create a Levantine and an Egyptian Arabic Treebank.

In the Arab world, the efforts are relatively limited in terms of creating annotated corpora. Examples include BZU's Curras, the Palestinian Arabic annotated corpus, NYUAD's Gumar, the Emirati Arabic annotated corpus, and Al-Mus'haf Quranic Arabic corpus. Another annotation effort with a focus on MSA spelling and grammar correction is the Qatar Arabic Language Bank (QALB), a project involving Columbia and CMUQ. Other specialized annotated corpora developed in the Arab world include NYUAD's parallel gender corpus with sentences in masculine and feminine for anti-gender bias research, the Arab-Acquis corpus pairing Arabic with all of Europe's languages for a portion of European parliamentary proceedings, and the MADAR corpus of parallel dialects created in collaboration with CMUQ.

In contrast to annotated corpora, there are many unannotated datasets. Most large datasets also started outside the Arab world, such as the Agence France Presse document collection, which is heavily used for Arabic information retrieval evaluation, the LDC's Arabic Gigaword, Arabic Wikipedia, and the ArTenTen corpus. Important collections in the Arab World include: the International Corpus of Arabic of Bibliotheca Alexandrina; Shamela, a large-scale corpus (1B words) covering the past 14 centuries

of Arabic; the Tashkeela corpus, containing 75M fully vocalized words (National Computer Science Engineering School in Algeria); NYUAD's Gumar Gulf Arabic corpus, containing over 100M words of Internet novels; and Abu El-Khair corpus (Umm Al-Qura University, Saudi Arabia). The success of word embedding models trained on unannotated data and resulting in improved performance for NLP tasks with little or no feature engineering has led to many contributions in Arabic NLP. The more recent appearance of contextualized embeddings trained on unannotated data, such as BERT, is creating promising possibilities for improving many Arabic NLP tasks. At the time of writing this article, a handful of contextualized embedding models are known to support Arabic including Multilingual BERT, AraBERT (AUB), GigaBert (Ohio State University), Marbert (University of British Columbia), and QARiB (QCRI).

Lexical resources. We can distinguish three types of lexical resources (that is, lexicons, dictionaries, and databases): morphological resources that encode all inflected forms of words; lexical resources that are lemma based, such as machine-readable monolingual and multilingual dictionaries; and semantic resources that link lemmas to each other, such as wordnets and ontologies. These resources are useful for a variety of NLP tasks.

Some of the earliest publicly available Arabic lexical resources were created outside of the Arab world in the second wave mentioned earlier. The Buckwalter Arabic Morphological Analyzer (BAMA), with its extended version called Standard Arabic Morphological Analyzer (SAMA), both available from the LDC, provided one of the first stem databases with tags and morphological solutions, and are used in a number of tools. Elixir-FM is a functional morphology analyzer developed at Charles University in the Czech Republic. The DIINAR Arabic morphological database is a full form resource developed in France. The Tharwa lemma-based lexicon was developed at Columbia University and included 70k entries in Egyptian Arabic, MSA, and English; it was later extended with Levantine Arabic.

Arabic WordNet is a semantic lexicon consisting of about 11k synsets, with subset and superset relationships between concepts, and linked to a number of other languages through the Global WordNet effort. This effort was done by a number of American and European universities. And the Arabic VerbNet classifies verbs that have the same syntactic descriptions and argument structure (University of Konstanz, Germany).

Some of the efforts in the Arab world led to multiple notable resources. Al-Khalil analyzer is a large morphological database for Arabic developed by researchers in Morocco and Qatar. Calima Star is an extension of the BAMA/SAMA family done at NYUAD and is part of the CAMEL Tools toolkit. BZU developed a large Arabic lexicographic database constructed from 150 lexicons that are diacritized and standardized. The MADAR project (NYUAD and CMUQ) includes a lexicon with 47k lemma entries of parallel statements in 25 city dialects. Other lexicons have been developed for Algerian, Tunisian, and Moroccan. Finally, in terms of semantic lexical resources, the BZU Arabic Ontology is a formal Arabic wordnet with more than 20k concepts that was built with ontological analysis in mind and is linked to the Arabic Lexicographic Database, Wikidata, and other resources.

More Arabic resources can be found in known international repositories (namely ELRA/ELDA, LDC, and CLARIN) or directly from their authors' websites.^{4,5,10} Unfortunately, many are not interoperable, have been built using different tools and assumptions, released under proprietary licenses, and few are comprehensive. Serious, well-planned, and well-coordinated investment in resources will be instrumental for the future of Arabic NLP.

Morphological processing. Given the challenges of Arabic morphological richness and ambiguity, morphological processing has received a lot of attention. The task of morphological analysis refers to the generation of all possible readings of a particular undiacritized word out of context. Morphological disambiguation is about identifying the correct in-



context reading. This broad definition allows us to think of word-level tasks such as part-of-speech (POS) tagging, stemming, diacritization, and tokenization as sub-types of morphological disambiguation that focus on specific aspects of ambiguity.

Most work on Arabic morphological analysis and disambiguation is on MSA; however, there is a growing number of efforts on dialectal Arabic. There are a number of commonly used morphological analyzers for Standard and dialectal Arabic (Egyptian and Gulf), including BAMA, SAMA, Elixir-FM, Al-Khalil, CALIMA Egyptian, and CALIMA Star. Some of the morphological disambiguation systems disambiguate the analyses that are produced by a morphological analyzer using PATB as a training corpus, for example, MADAMIRA (initially developed at Columbia University) and other variants of it from NYUAD. Farasa (from QCRI) uses independent models for tokenization and POS tagging.

Syntactic processing. Syntactic parsing is the process of generating a parse tree representation for a sentence that indicates the relationship among its words. For example, a syntactic parse of the sentence أَرَقَ اُتْدِي دَجَلْ اَبَاتِكْفَلْ اُنْبِلْ اَطْلْ ا [lit.] *read the-student the-book the-new*; 'the student read the new book,' would indicate that the adjective *the-new* modifies the noun *the-book*, which itself is the direct object of the verb *read*.

There are many syntactic representations. Most commonly used in Arabic are the PATB constituency representation, the CATiB dependency representation, and the Universal Dependency (UD) representation. All of these were developed outside of the Arab world. The UD representation is an international effort, where NYUAD is the representative of the Arab world on Arabic.

The most popular syntactic parsers for Arabic are: Stanford, Farasa (QCRI), and CamelParser (NYUAD). Stanford is a statistical parser from the Stanford Natural Language Processing Group that can parse English, German, Arabic, and Chinese. For Arabic, it uses a probabilistic context free grammar that was developed based on PATB. Farasa is an Arabic NLP toolkit that provides syntactic constituency and dependency parsing. CamelParser is a dependency parser trained on CATiB treebank using MaltParser, a language-independent and data-driven dependency parser. A discussion and survey of some of the Arabic parsing work is presented in Habash.⁵

Named entity recognition (NER) is the task of identifying one or more consecutive words in text that refer to objects that exist in the real world (named entities), such as organizations, persons, locations, brands, products, foods, and so forth. NER is essential for extracting structured data from an unstructured text, relationship extraction, ontology



population, classification, machine translation, question answering, and other applications. Among the challenges facing Arabic NER compared to English NER is the lack of letter casing, which strongly helps English NER, and the high degree of ambiguity, including especially confusable proper names and adjectives, for example, *كاريوم* *kariym* can be the name 'Kareem' or the adjective 'generous.'

Arabic NER approaches include the use of hand-crafted heuristics, machine learning, and hybrids of both with heavy reliance on gazetteers. Much of the earlier work on Arabic NER focused on formal text, typically written in MSA. However, applying models trained on MSA text to social media (mostly dialectal) text has led to unsatisfactory results. Recent contextualized embeddings and other deep learning approaches such as sequence-to-sequence models and convolutional neural networks have led to improved results for Arabic NER. As with other utilities, early research was done outside of the Arab world, but more work is now happening in the Arab world. An extensive list of Arabic NER challenges and solutions can be found in Shaalan.⁹

Dialect identification (DID) is the task of automatically identifying the dialect of a particular segment of speech or text of any size: word, sentence, or document. This task has been attracting increasing attention in NLP for a number of language

varieties. DID has been shown to be important for several NLP tasks where prior knowledge about the dialect of an input text can be helpful, such as machine translation, sentiment analysis, and author profiling.

Early Arabic multi-dialectal data sets and models focused on the regional level. The Multi Arabic Dialects Application and Resources (MADAR) project aimed to create a finer grained dialectal corpus and lexicon. The data was used for dialectal identification at the city level of 25 Arab cities, and was used in a shared task for DID. The main issue with that data is that it was commissioned and not naturally occurring. Concurrently, larger Twitter-based datasets covering 10-to-21 countries were also introduced. The Nuanced Arabic Dialect Identification (NADI) shared task followed earlier pioneering works by providing country-level dialect data for 21 Arab countries, and introduced a province-level identification task aiming at exploring a total of 100 provinces across these countries. Earlier efforts started in the west, most notably in Johns Hopkins University, but more work is happening now in the Arab world, at NYUAD and QCRI, for example.

Infrastructure. To aid the development of NLP systems, a number of multi-lingual infrastructure toolkits have been developed, such as GATE,^d

d <https://gate.ac.uk>

Stanford CoreNLP,^e and UIMA.^f They offer researchers easy access to several tools through command-line interfaces (CLIs) and application programming interfaces (APIs), thus eliminating the need to develop them from scratch every time. While Arabic NLP has made significant progress with the development of several enabling tools, such as POS taggers, morphological analyzers, text classifiers, and syntactic parsers, there is a limited number of homogeneous and flexible Arabic infrastructure toolkits that gather these components. MADAMIRA is a Java-based system providing solutions to fundamental NLP tasks for Standard and Egyptian Arabic. These tasks include diacritization, lemmatization, morphological analysis and disambiguation, POS tagging, stemming, glossing, (configurable) tokenization, base-phrase chunking, and NER.^g Farasa^h is a collection of Java libraries and CLIs for MSA. These include separate tools for diacritization, segmentation, POS tagging, parsing, and NER. SAFARⁱ is a Java-based framework bringing together all layers of Arabic NLP: resources, pre-processing, morphology, syntax, and semantics. CAMEL Tools is a recently developed collection of open source tools, developed in Python, that supports both MSA and Arabic dialects.^j It currently provides APIs and CLIs for pre-processing, morphological modeling, dialect identification, NER, and sentiment analysis. Other notable efforts include AraNLP, ArabiTools,^k and Adawat.^l A feature comparison of some Arabic infrastructures can be found in Obeid,⁸ while a detailed survey and a software engineering comparative study can be found in Jaafar.⁶

Arabic NLP Applications

Machine translation (MT) is one of the earliest and most worked on areas in NLP. The task is to map input text in a source language such as English

e <https://stanfordnlp.github.io/CoreNLP/>

f <https://uima.apache.org/d/uimaj-current/>

g <https://camel.abudhabi.nyu.edu/madamira/>

h <https://farasa.qcri.org/>

i <http://arabic.emi.ac.ma/safar/>

j <https://github.com/CAMEL-Lab>

k <https://www.arabitools.com/>

l <http://adawat.sourceforge.net/>

to an output text in a target language such as Arabic. Early MT research was heavily rule-based; however, now it is almost completely corpus-based using a range of statistical and deep learning models, depending on resource availability.

For MSA, parallel data in the news domain is plentiful.^m There are other large Arabic parallel collections under the OPUS project and as part of the United Nations' six-language parallel corpus. Other specialized corpora include the Arab-Acquis corpus pairing with European languages developed in NYUAD, and the AMARA educational domain parallel corpus developed by QCRI. Dialectal parallel data are harder to come by and most are commissioned translations.

There are many other efforts in Statistical MT (SMT) from and to Arabic. Recently, deep neural networks have been adopted for Arabic machine translation. While most researched MT systems for Arabic target English, there have been efforts on MT for Arabic and other languages, including Chinese, Russian, Japanese, and all of the European Union languages.

MT for Arabic dialects is more difficult due to limited resources, but there are noteworthy efforts exploiting similarities between MSA and dialects in universities and research group around the world. Finally, there is a notable effort on Arabic sign-language translation at King Fahd University of Petroleum and Minerals. For recent surveys of Arabic MT, see Ameer.¹ Despite all these contributions, much research work is still needed to improve the performance of machine translation for Arabic.

Pedagogical applications (PA) focus on building tools to develop or model for four major skills: reading, writing, listening, and speaking. Arabic PA research has solely focused on MSA. PA systems can be distinguished in terms of their target learners as first language (L1) or second (foreign) language (L2) systems. This distinction can be problematic since, for Arabs, learning to read MSA is somewhat akin to reading a foreign

tongue due to its lexical and syntactic divergence from native dialects. We focus our Arabic PA discussion on (a) computer-assisted language learning (CALL) systems, (b) readability assessment, and (c) resource-building efforts.


CALL systems utilize NLP enabling technologies to assist language learners. There has been a number of efforts in Arabic CALL exploring a range of resources and techniques. Examples include the use of Arabic grammar and linguistic analysis rules to help learners identify and correct a variety of errors; and multi-agent tutoring systems that simulate the instructor, the student, the learning strategy, and include a logbook to monitor progress, and a learning interface. Another approach focuses on enriching the reading experience with concordances, text-to-speech, morpho-syntactic analysis, and auto-generated quiz questions.

Readability assessment is the task of automatic identification of a text's readability, that is, its ability to be read and understood by its reader employing an acceptable amount of time and effort. There has been a range of approaches for Arabic L1 and L2 readability. On one end, we find formulas using language-independent variables such as text length, average word length, and average sentence length, number of syllables in words, the relative rarity or absence of dialectal alternatives, and the presence of less common letters. Others integrate Arabic morphological, lexical, and syntactic features with supervised machine learning approaches.

Although some progress has been made for both L1 and L2 PA, the dearth of resources compared with English remains the bottleneck for future progress. Resource-building efforts have focused on L1 readers with particular emphasis on grade school curricula. There is a push to inform the enhancement of curricula using pedagogical tools and to compare curricula across Arab countries. The L2 PAs are even more constrained, with limited corpora and a disproportionate focus on beginners.ⁿ There is a definite need



Current NLP methods for Arabic language dialogue are mostly based on handcrafted rule-based systems and methods that use feature engineering.



^m Linguistic Data Consortium (LDC) resources: LDC2004T18, LDC2004T14, and LDC2007T08.

ⁿ <https://learning.aljazeera.net/en>

It is time to have an association for Arabic language technologists that brings together talent and resources and sets standards for the Arabic NLP community.

for augmenting these corpora in a reasoned way, taking into consideration different text features and learners, both young and old, beefing up the sparsely populated levels with authentic material, and exploiting technologies such as text simplification and text error analysis and correction. Learner corpora, which as the name suggests are produced by learners of Arabic, can inform the creation of tools and corpora. A recent effort developed a large-scale Arabic readability lexicon compatible with an existing morphological analysis system.

Information retrieval and question answering. With the increasing volume of Arabic content, information retrieval, or search, has become a necessity for many domains, such as medical records, digital libraries, web content, and news. The main research interests have focused on retrieval of formal language, mostly in the news domain, with ad hoc retrieval, OCR document retrieval, and cross-language retrieval. The literature on other aspects of retrieval continues to be sparse or non-existent, though some of these aspects have been investigated by industry. Other aspects of Arabic retrieval that have received some attention include document image retrieval, speech search, social media and web search, and filtering.³ However, efforts on different aspects of Arabic retrieval continue to be deficient and severely lag behind efforts in other languages. Examples of unexplored problems include searching Wikipedia, which contains semi-structured content, religious texts, which often contain semi-structured data such as chains of narrations, rulings, and commentaries, Arabic forums, which are very popular in the Arab world and constitute a significant portion of the Arabic Web, and poetry. To properly develop algorithms and methods to retrieve such content, standard test sets and clear usage scenarios are required. We expect that recent improvements in contextual embeddings can positively impact the effectiveness of many retrieval tasks.

Another information retrieval-related problem is question answering, which comes in many flavors, the

most common of which is attempting to identify a passage or a sentence that answers a question. Performing such a task may employ a large set of NLP tools such as parsing, NER, coreference resolution, and text semantic representation. There has been limited research on this problem, and existing commercial solutions such as Ujeeb.com are rudimentary.

Dialogue Systems

Automated dialog systems capable of sustaining a smooth and natural conversation with users have attracted considerable interest from both research and industry in the past few years. This technology is changing how companies engage with their customers among many other applications. While commercial dialog systems by big multinational companies such as Amazon's Alexa, Google's Home, and Apple's Siri support many languages, only Apple's Siri supports Arabic with limited performance. There are some strong recent competitors in the Arab world, particularly Arabot^o and Mawdoo3's Salma.^p

While there is an important growing body of research on English language dialog systems, current NLP methods for Arabic language dialogue are mostly based on handcrafted rule-based systems and methods that use feature engineering. Among the earliest research efforts on Arabic dialog applications is the Quran chatbot, where the conversation length is short since the system answers a user input with a single response. It uses a retrieval-based model as the dataset is limited by the content of the Quran. A recent approach used deep learning techniques for text classification and NER to build a natural language understanding module—the core component of any dialogue system—for the domain of home automation in Arabic. A unique dialogue system from NYUAD explored bilingual interfaces where Arabic speech can be used as input to an English bot that displays Arabic subtitles. Other works have focused on developing dialog systems for the case of Arabic dialects, as with the publicly avail-

^o <https://arabot.io/>

^p <http://salma.ai/>

able NYUAD Egyptian dialect chatbot *Botta*, and KSU's Saudi dialect information technology-focused chatbot *Nabiha*.

Sentiment and Emotion Analysis

Sentiment analysis (SA), or opinion mining, is the task of identifying the affective states and subjective information in a text. For example, an Egyptian Arabic movie review such as *أيد تهنسلا م لي ف نسحأ* 'the best movie this year!' is said to indicate a positive sentiment. SA is a very powerful tool for tracking customer satisfaction, carrying out competition analysis, and generally gauging public opinion towards a specific issue, topic, or product. SA has attracted a lot of attention in the Arabic research community during the last decade, connected with the availability of large volumes of opinionated and sentiment reflecting data from Arabic social media. Early Arabic SA efforts focused on the creation of needed resources such as sentiment lexicons, training datasets, and sentiment treebanks, as well as shared task benchmarks. Arabic SA solutions span a range of methods from the now conventional use of rules and lexicons to machine learning based methods as well as hybrid approaches employing morphological and syntactic features. Recently, fine-tuning large pre-trained language models has achieved improved Arabic SA results. Arabic emotion detection is a closely related topic that has attracted some attention recently. It aims to identify a variety of emotions in text such as anger, disgust, surprise, and joy. Similar to how SA resources and models started maturing, a lot of work still needs to be done in emotion detection. Another related problem is stance detection, which attempts to identify positions expressed on a topic or towards an entity. Stances are often expressed using non-sentiment words. For a recent comprehensive survey on the status of Arabic SA and the future directions, see Badaro et al.²


Content Moderation on Social Media

The task of content moderation is about the enforcement of online outlets' policies against posting user comments that contain offensive

language, hate speech, cyber-bullying, and spam, among other types of inappropriate or dangerous content.^q Such content cannot be easily detected given the huge volume of posts, dialectal variations, creative spelling on social media, and the scarcity of available data and detection tools. This area is relatively new for Arabic. One of the more active areas has to do with the detection of offensive language, which covers targeted attacks, vulgar and pornographic language, and hate speech. Initial work was performed on comments from a news site and a limited number of tweets and YouTube comments. Some works focused on adult content and others on hate speech. Recent benchmarking shared tasks included the automatic detection of such language on Twitter. Work on spam detection on Twitter is nascent and much work is required.

Future Outlook

Arabic NLP has many challenges, but it has also seen many successes and developments over the last 40 years. We are optimistic by its continuously positive albeit sometimes slow development trajectory. For the next decade or two, we expect a large growth in the Arabic NLP market. This is consistent with global rising demands and expectations for language technologies and the increase in NLP research and development in the Arab world. The growing number of researchers and developers working on NLP in the Arab world makes it a very fertile ground ready for major breakthroughs. To support this vision, we believe it is time to have an association for Arabic language technologists that brings together talent and resources and sets standards for the Arabic NLP community. Such an organization can support NLP education in the Arab world, serve as a hub for resources, and advocate for educators and researchers in changing old-fashioned university policies regarding journal-focused evaluation, and encouraging collaborations within the Arab world by connecting academic, industry, and government-

tal stakeholders. We also recommend more open source tools and public data be made available to create a basic development framework that lowers the threshold for joining the community, thus attracting more talent that will form the base of the next generation of Arabic NLP researchers, developers, and entrepreneurs. 

References

1. Ameur, M.S.H., Meziane, F., Guessoum, A. Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Rev.* 38 (2020), 100305.
2. Badaro, G., et al. A survey of opinion mining in Arabic: A comprehensive system perspective covering challenges and advances in tools, resources, models, applications, and visualizations. *ACM Trans. Asian and Low-Resource Language Information Processing (TALLIP)* 18, 3 (2019), 1–52.
3. Darwish, K., Magdy, W. Arabic information retrieval. *Foundations and Trends in Information Retrieval* 7, 4 (2014), 239–342.
4. Farghaly, A., Shaalan, K. Arabic natural language processing: Challenges and solutions. *ACM Trans. Asian Language Information Processing* 8, 4 (2009), 1–22.
5. Habash, N.Y. *Introduction to Arabic Natural Language Processing*, Vol. 3. Morgan & Claypool Publishers, 2010.
6. Jaafar, Y., Bouzoubaa, K. A survey and comparative study of Arabic NLP architectures. *Intelligent Natural Language Processing: Trends and Applications*. Springer, 2018, 585–610.
7. Mubarak, H., et al. Overview of OSACT4 Arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, May 2020. European Language Resource Association, 48–52.
8. Obeid, O., et al. CAMEL tools: An open source Python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conf.*, May 2020, European Language Resources Association, 7022–7032.
9. Shaalan, K. A survey of Arabic named entity recognition and classification. *Computational Linguistics* 40 (2014), 469–510.
10. Zaghoulani, W. Critical survey of the freely available Arabic corpora. In *Proceedings of the Workshop on Open-Source Arabic Corpora and Processing Tools* (2014), 1–8.

Kareem Darwish, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.

Nizar Habash, New York University Abu Dhabi, United Arab Emirates.

Mourad Abbas, Center of Scientific and Technical Research for the Development of Arabic Language (CRSTDLA), Bouzareah, Algeria.

Hend Al-Khalifa, King Saud University, Riyadh, Saudi Arabia.

Huseein T. Al-Natsheh, Mawdoo3, Jordan.

Houda Bouamor, Carnegie Mellon University, Doha, Qatar.

Karim Bouzoubaa, Mohammed V University, Rabat, Morocco.

Violetta Cavalli-Sforza, Al Akhawayn University, Ifrane, Morocco.

Samhara R. El-Beltagy, Newgiza University, Cairo, Egypt.

Wassim El-Hajj, American University of Beirut, Beirut, Lebanon.

Mustafa Jarrar, Birzeit University, Palestine.

Hamdy Mubarak, Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar.

Copyright held by authors/owners.
Publication rights licensed to ACM.

^q <https://www.bbc.co.uk/usingthebbc/terms/what-are-the-rules-for-commenting/>

BY DAVID KEYES

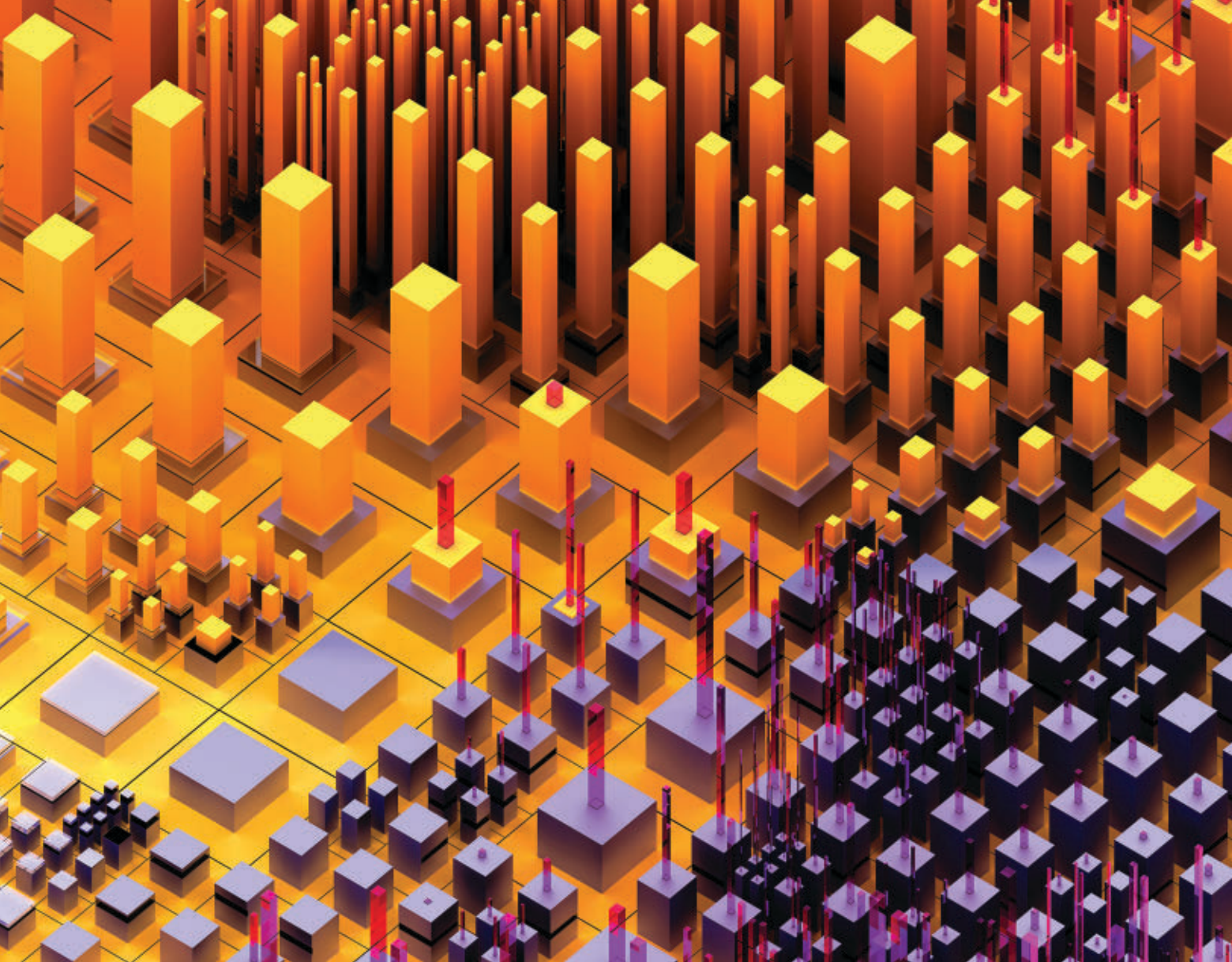
The Arab World Prepares the Exascale Workforce

THE ARAB WORLD is currently host to eight supercomputers in the Top500 globally, including the current #10 and a former #7. Hardware can become a honeypot for talent attraction—senior talent from abroad, and rising talent from within. Good return on investment from leading-edge hardware motivates forging collaborative ties to global supercomputing leaders, which leads to integration into the global campaigns that supercomputing excels in, such as predicting climate change and developing sustainable energy resources for its mitigation, positing properties of new materials and catalysis by design, repurposing already-certified drugs and discovering new ones, and big data analytics and machine learning applied to science and to society. While the petroleum industry has been the historical motivation for supercomputing in the Arab World,

with its workloads of seismic imaging and reservoir modeling, the attraction today is universal.

However, it is not sufficient to install and boot supercomputers. Their purpose is performance, and their acquisition and operating costs are too high to use them any other way. In each phase of computation, the limiting resource must be identified and computation reorganized to push the bottleneck further away, ideally guided by a performance model. The soul of the machine is the software: the distributed shared memory data structures, the task graphs, the communication patterns. The software is generally not performance-portable; it must be re-tuned in each application-architecture context. As applications become more ambitious and architectures become more austere, algorithms and software must bridge the growing gap.

Hands-on opportunities to resolve this application-architecture tension are a lure to students who had not previously considered supercomputing careers. In some cases, they must surmount significant hurdles in mathematical or computational preparation to enlist, but enlist they do, and they often wind up in globally leading institutions upon graduation. The stories in this article grew up around a university-operated supercomputer, but they largely can be replicated with much less investment because the main challenges today are not in coordinating tens of thousands of nodes across a low-latency, high-bandwidth network. Rather, the challenges lie in extracting performance *from within* increasingly heterogeneous nodes. Furthermore, the cloud now provides high-performance computing (HPC) environments. It is estimated that the percentage of HPC jobs run in the public cloud nearly doubled from 10%–12% in 2018 to 20% in 2019.⁴ Many supercomputers, including the currently #1-ranked Fugaku (featuring ARM-based Fujitsu A64fx) and #2-ranked Summit (featuring



NVIDIA V100 and IBM Power9), offer small competitively awarded research accounts at no cost. (Disclosure: the author's team presently employs accounts on each.) We argue that a menu of *Awareness, Examples, Instruction, Opportunity, and Utilization* will Yield members of the exascale workforce, mnemonically: *A, E, I, O, U* and sometimes—hopefully often(!)—*Y*.

From the Middle East to the Best of the West

The basis for confidence in this five-fold agenda for preparing students for the heterogeneous environments of exascale computing is empirical: nine of the author's earliest Ph.D. students from the King Abdullah University of Science and Technology (KAUST, founded in 2009 in Saudi Arabia) received as their first job offer a U.S. DOE-funded post-doc, either in NERSC's NESAP, NNSA's PSAAP, or

the agency-wide ECP. All of them held their U.S. DOE-based job offer before they defended their dissertations.

Three Saudi Ph.D.s beyond the nine sought by the U.S. DOE elected to stay at home and grow their careers with their increasingly information-based national economy. One joined Boeing and another joined NEOM, the futuristic green city along the Gulf of Aqaba billed as “an accelerator of human progress” that has 10 times the land area of Hong Kong. The last joined a digital start-up he had co-founded on the side as a student, already with Series A financing of over USD \$2 million. Eleven of these 12 students hail from the Arab World: from Egypt, Jordan, Lebanon, Saudi Arabia, and Syria. Some students had their software integrated into NVIDIA's cuBLAS or Cray's LibSci before they graduated, and one had their software integrated into a prototype of Saudi Aramco's

next-generation seismic inversion code; integration into NEC's Numeric Library Collection is pending.

Of these 12 HPC Ph.D.s, four are women. All but two completed their bachelor's degrees in Arab World universities in departments of computer science, computer engineering, or information science. Only one was proactively recruited into an HPC-oriented fellowship at KAUST. The others came from a globally cast net offering doctoral fellowships to study computer science or applied mathematics more generally and were lured to supercomputing by its opportunity.

Two of the 2020 graduates won major conference awards for papers based on their thesis work: Noha Alharthi lead-authored *Solving Acoustic Boundary Integral Equations using High Performance Tile Low-Rank LU Factorization*, which was awarded the Gauss Center for Supercomputing Award in

June 2020 at the (virtualized) 35th International Supercomputing Conference (ISC'20), and Tariq Alturkestani lead-authored *Maximizing I/O Bandwidth for Reverse Time Migration on Heterogeneous Large-Scale Systems*, which was awarded Best Paper at the (virtualized) 26th Euro-Par Conference in September 2020. Each paper is interdisciplinary: Alharthi's spans the discretization of singular integral equations, massively distributed data-sparse linear algebra, and acoustic scattering from irregularly shaped bodies, while Alturkestani's generalizes 2-level cache protocols to N -level memory hierarchies for hiding pre-fetching and write-back times of the huge datasets of reverse-time migration on massively distributed systems (including globally #2-ranked supercomputer Summit) for seismic imaging of petroleum deposits. Topics pursued by the graduates who were recruited to U.S. DOE-funded post-docs include: dense and hierarchically low-rank linear algebra kernels implemented on graphics processing units (GPUs), singular value decomposition (SVD) and eigensolvers implemented on massively distributed memory systems; high-order stencil update protocols for Cartesian lattices implemented on many-core central processing units (CPUs) with shared caches; a many-core implementation of an unstructured-grid implicit external aerodynamics code, and a

fast-multipole method preconditioned boundary-integral equation solver.

Following their DOE-sponsored post-docs, two of the U.S.-based alumni moved to Intel, one to NVIDIA, and one to a machine learning start-up in the Bay Area; the rest remain in DOE Exascale Computing Project (ECP)-funded research positions. Not all of them dealt with heterogeneity as a first-class consideration for their theses, but many implementations were GPU or hybrid. While all scaled in a performance-oriented way to distributed memory, the main contributions took place within a node. They were immersed in a roofline modeling culture and became fluent with DAG-based dynamic runtime programming. Hands-on, they compared as many vendor platforms as were available locally and then gained access to guest accounts abroad by sharing tantalizing locally generated results. They also compared against the prior state of the art on a given platform, whether libraries like MKL, CuBLAS, PLASMA, or MAGMA, or runtimes like ParSEC, QUARK, StarPU, or OpenMP-LLVM. All but a couple released their codes at github.com/ecrc, where the visibility of one code calls attention to another.

Facing the 'Universals' of Exascale Computing

As part of their inauguration into research, each student was presented

with a list of "universals" for exascale computing that cuts across most applications. The list has been growing over the past few years to include the 15 grouped in Figure 1 into five architectural imperatives, five strategies already widely in practice, and five strategies in progress. Each student was asked to identify a research contribution among these "universals" and to adopt a particular demanding application to keep the work practically motivated, typically through a co-advisor. From our experience with this cadre of students, we advocate the following five principles for equipping the next-generation exascale workforce.

Awareness of heterogeneity should be emphasized from the beginning. Heterogeneity of processing, memory, and network elements is now the norm, driven by opportunities for energy efficiency for specialized instructions, such as the $D \leftarrow A * B + C$ 4x4 matrix-multiply-and-add in convolutional neural networks that do 64 FMADD operations in one instruction. Increasingly, performance-oriented programmers will make choices about which memories to stash their data structures in, and how to route their data, possibly doing operations like transposes or reductions of the data *en route*. In Figure 2, we adapt a figure from a DOE report on exascale architecture¹ by adding deep learning, quantum, and neuromorphic elements. While a quantum device likely will need to be off-board for cryogenic engineering purposes, students should, for example, recognize that an unconstrained optimization step in a large scientific code may be ideal to offload to such a device in the future to quickly examine billions of random possibilities and return one that, with high probability, is within a tight tolerance of the optimum. Programmers of the exascale era have to think about *what runs best where*.

Examples should be provided. Students should read success stories about applications that profit from heterogeneity and how; for example, ACM Gordon Bell Prize finalist papers, like the 2019 OMEN code with its data-centric DaCE programming model,⁵ or the 2020 DeePMD-kit with its use of machine learning to replace inner loops of expensive floating-

Figure 1. Traditional conditional structure.

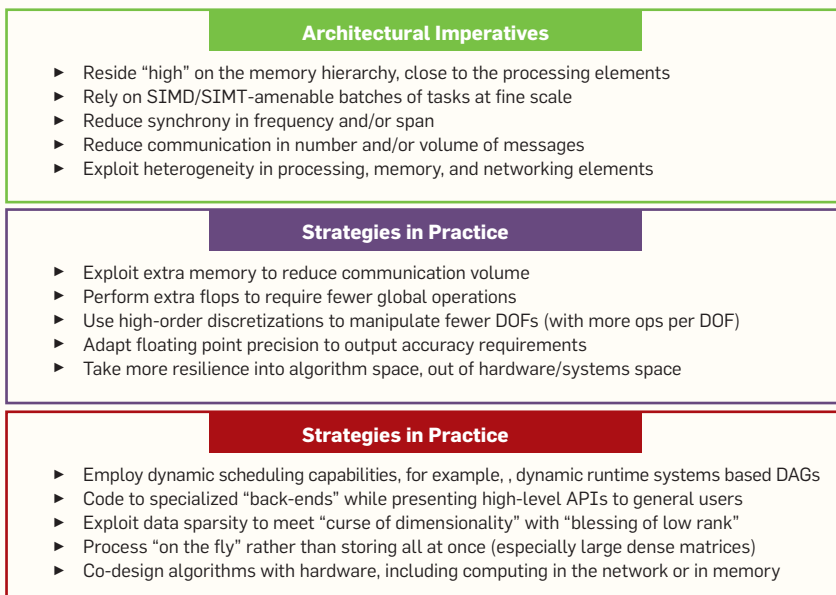
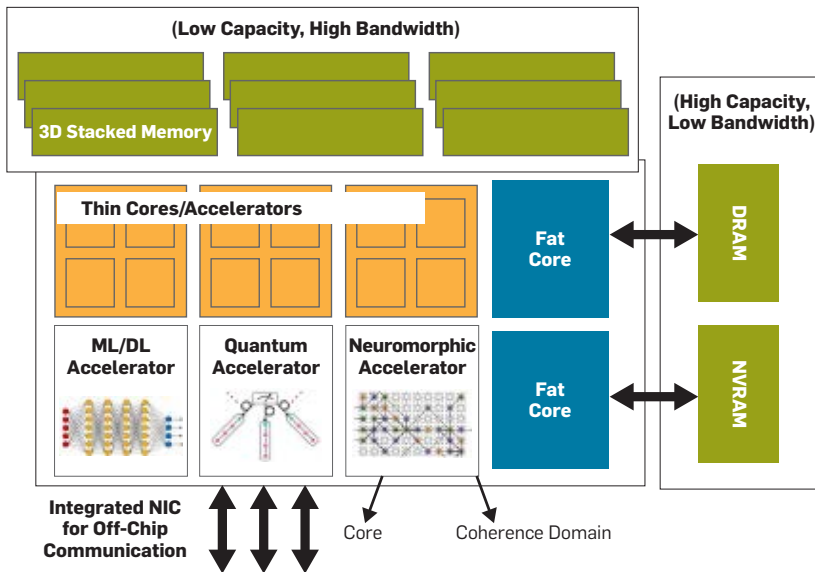


Figure 2. Augmented example of a heterogeneous node from Ang et al.¹

The soul of the machine is the software: the distributed shared memory data structures, the task graphs, the communication patterns.

point function evaluations of *ab initio* electronic structure calculations.³ Today, the examples employ vector units, GPUs, tensor processing units (TPUs), and field-programmable gate arrays (FPGAs); in mainstream scientific campaigns, more neuromorphic and quantum devices may soon be relevant. Memory systems stretch from registers, through multiple levels of cache with varying nestedness, to HBM, DRAM, NVRAM, local disk, and federated data bases. Communication channels range from direct optical, through copper, to optical fiber. For petabyte datasets, users need to consider whether they should use a courier service, or leave data globally distributed and manage it as a federated entity.

Instruction should be given on two levels. High-level multidisciplinary thinking estimates thresholds for using a technique that amortize the overheads, and how to recognize amenable kernels in applications. Low-level training in how to express scheduling, data placement, and heterogeneous targets such as vector extensions, CUDA, and libraries for remote operations, is also important. Syntax often can be taught outside of credit-bearing courses, such as through hosting vendors for weekend tutorial/hackathons, while the conceptual parts belong in proper courses.

Opportunity to experience develop-

ment at the cutting edge motivates and equips. Sometimes, this is best accomplished in a three-month to six-month internship. The students described here interned at mission-oriented research labs like Argonne; academic institutions with performance expertise we needed like Erlangen, home of the LikWid performance tools,^a HPC vendors like NVIDIA, and HPC customers like Saudi Aramco. In some cases, the thesis topic arose from the internship advisor. In other cases, the application that motivated the algorithmic innovation or implementation of the thesis was mastered in the internship.

Utilization is the ultimate goal: hands-on code development as part of the thesis, ideally plugging into a multidisciplinary team so the specialist effort is part of something bigger and real-world, like clean combustion, vehicle aerodynamics, or geospatial statistical inference of the weather. The application motivates, may lead to sponsorship, and brings visibility beyond the algorithmic and software accomplishments.

Fashioning a Computational Mecca

Without question, the lure of top Arab World talent to the opportunities of exascale computing that are being established beyond *and within* the

^a <https://github.com/RRZE-HPC/likwid>.

The function of the ECRC is best understood in terms of the ‘hourglass model’ of software, a concept borrowed from the TCP-IP philosophy.

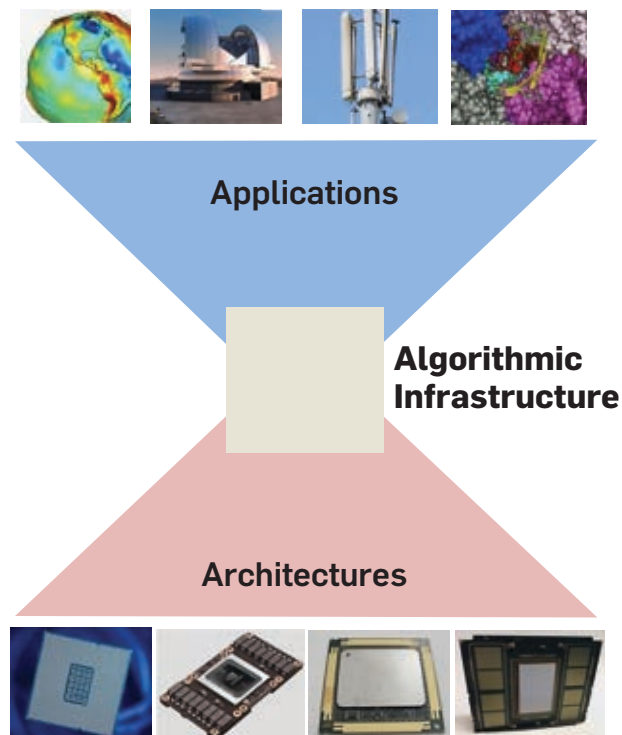
Arab World itself is due in part to access to a petascale supercomputer as a technological stepping stone and a sheer source of inspiration. However, universities need not have the means to bring a supercomputer to their campus to participate. Besides access to HPC in the commercial cloud (which shifts the expense from capital to operating), many of the more than 500 petascale supercomputers in the world in the hands of universities or national laboratories offer exploratory grants at no cost, including to off-shore collaborators, and some also offer summer training programs in hopes of attracting a future workforce. More importantly, as mentioned earlier, the most significant bottlenecks to performance scalability now lie within the individual (often heterogeneous) nodes, meaning that a modest collection of the latest processor offerings from providers such as AMD, ARM, Fujitsu, IBM, Intel, and NVIDIA put experimental proofs of concept within reach.

The Extreme Computing Research Center (ECRC), which sponsored the students discussed earlier, was created outside of KAUST’s 16 degree

programs as one of 14 mission-oriented research centers. The Centers create critical mass beyond the capacity of individual faculty members to encourage translating basic research into translational ends. They induce faculty and students from the degree programs to their missions with expertise, centrally supplied competitively awarded funding, space, research facilities, and reputation. The Centers support a small number of experienced research scientists. In the case of the ECRC, these are professional software engineers who, together with the faculty, contribute beyond the university to such industry-standard open-source libraries as PETSc, MFEM, SPECfem3D, ug4, CLAWPACK and pyCLAW, mpi4py, and OpenFOAM.

The function of a center dedicated to software infrastructure is best understood in terms of the “hourglass model” of software, in Figure 3, a concept borrowed from the TCP-IP philosophy:² many diverse scientific applications (the top of the hourglass) are enabled to run with high performance on many diverse computer architectures (the bottom of the hourglass)

Figure 3. An hourglass model for scientific software.



through a standard interface (the neck of the hourglass) implemented as callable software libraries whose purpose is to partially hide the complexity and diversity of the architecture. The role of such a center becomes more interesting as architectures evolve under the premium of energy efficiency to become more specialized to certain tasks, thus presenting a host of heterogeneous resources in processor, memory, and network components. The diversity of applications here refers both to domain subject matter, from seismic imaging to genome-wide association studies, and to technique, from simulation based on first-principles models to machine learning, where first-principles models are not readily constructed but input-output maps can be learned from data.

There are manifold ways to improve a scientific computation, such as: increase its *accuracy* (computational resolution of an underlying continuum); increase its *fidelity* (inclusion of a system's full features in a computational model); tighten its *uncertainty* (bound the error of a model's output in terms of errors in its inputs); and reduce its *complexity* (computational costs, in terms of storage and operations) to achieve a sought accuracy, fidelity, and confidence. Modelers generally customize the first three to their application and are happy to hand off the fourth, complexity reduction and architectural tuning, as a productive separation of concerns. KAUST's ECRC focuses its resources on complexity reduction and architectural tuning for widely used computational kernels in simulation and data analytics.

Ph.D. students can become a source of widely distributed software for the implementation of efficient algorithms for simulation and data analytics on high-performance hardware by facilitating the transition to algorithms that exploit a hierarchy of scales, such as multigrid, fast multiple, hierarchical low-rank matrices, and hierarchical coarsenings of graphs. Hierarchical algorithms are much more efficient than their traditional "brute force" counterparts, but also more complex to implement because of their nonuniformity of scales. These algorithms exploit

the mathematics of "data sparsity," architecture-specific instruction-level concurrency, and they aim to reduce communication and synchronization, the latter much more expensive than operations on cached data. As a further step, algorithms that exploit randomization, such as stochastic gradient descent in machine learning and algorithms based on randomized subspace selection in linear algebra, can to *high probability* deliver a *highly accurate* answer at a much *lower cost* than their deterministic equivalents.

Technology translation for software includes both computer vendors (for example, Cray, NVIDIA) and commercial users (such as Aramco, McLaren). Translation efforts train post-docs and master's students, as well as the Ph.D. students emphasized herein, for the rapidly expanding workforce in simulation and big data analytics, for placement in the world's leading computing establishments for the future of the national economy, such as at Saudi Aramco, which operates three of the world's Top500 systems. Some ECRC members also carry out computational science and engineering campaigns of their own.

The ECRC vision fits in the "digital pillar" of KAUST's 2020–2025 strategic plan (see the article by Elmootazbellah Elnozahy in this special section), especially with respect to climate prediction and artificial intelligence, with also a recent foray into smart health. Traditionally, it has supported other institutional pillars, especially energy and environment. More than half (90 as of the time of this writing) of KAUST's faculty have accounts on KAUST's supercomputer Shaheen-2, currently the third most-powerful system in the Arab World and one of the few most powerful operated by any university on behalf of its own researchers. Eighteen research organizations beyond KAUST in Saudi Arabia have accounts on Shaheen-2. Tellingly, many of these users are KAUST alumni who now work in ministries or other universities. They were the first to bring expectations from supercomputing into their organizations. This illustrates the "flying embers" effect of a supercomputer.

Student theses in such innovations

as data sparsity in linear algebra on GPUs through its HiCMA and H2Opus software—a critical technology in spatial statistics and engineering optimization, and rapid mesh traversal on many-core shared-memory accelerators through its GIRIH software—a critical technology in seismic wave propagation, have ultimately attracted collaborations with major vendors to the Arab World. Today, the majority of the members of the exascale workforce trained in the Arab World find their best opportunities in countries already possessing a fully developed supercomputing ecosystem, including more capable supercomputer hardware. With the UAE and Morocco recently joining Saudi Arabia in operating a Top 100 supercomputer and with increasing trends in simulation and analytics/machine learning in all disciplines of science and engineering, we expect an increasing fraction of the workforce trained in Saudi Arabia will remain in Saudi Arabia and become the human core of the regional ecosystem. For a decade now, Saudi Arabia has hosted a High-Performance Computing Symposium that attracts researchers from around the region and allows students to mix in poster sessions and sense the regional spirit of HPC. In becoming a source of students and a source of software, a university can aspire to become a "mecca" for high-performance computing. 

References

1. Ang, J.A. (Ed) et al. Abstract machine models and proxy architectures for exascale computing. In *Proceedings of Hardware-Software Co-Design for High-Performance Computing*, (New Orleans, LA, 2014), 25–32; doi: 10.1109/Co-HPC.2014.4.
2. Beck, M. On the hourglass model. *Commun. ACM* 62, 7 (June 2019), 48–57.
3. Jia, W., Wang, H., Chenz, M., Luz, D., Lin, L., Car, R., Weinan E. and Zhang, L. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *Proceedings of the Intern. Conf. High-Performance Computing, Networking, Storage and Analysis*, (Nov. 2020).
4. Norton, A., Conway, S. and Joseph, E. *Bringing HPC Expertise to Cloud Computing*. Opinion Whitepaper, Hyperion Research, Apr. 2020.
5. Ziogas, A.N., Ben-Nun, T., Fernández, G.I., Schneider, T., Luisier, M. and Hoefler, T. A data-centric approach to extreme-scale Ab initio dissipative quantum transport simulations. In *Proceedings of the Intern. Conf. High-Performance Computing, Networking, Storage and Analysis*, (Nov. 2019).

David Keyes is a professor of applied mathematics and computational science and director of the Extreme Computing Research Center at the King Abdullah University of Science and Technology, Saudi Arabia.



BY INGMAR WEBER, MUHAMMAD IMRAN,
FERDA OFLI, FOUAD MRAD, JENNIFER COLVILLE,
MEHDI FATHALLAH, ALISSAR CHAKER,
AND WIGDAN SEED AHMED

Non-Traditional Data Sources: Providing Insights into Sustainable Development

THE WORLD IS facing enormous challenges, ranging from climate change to extreme poverty. The 2030 Agenda for Sustainable Development and its 17 Sustainable Development Goals (SDGs)^a were adopted by United Nations Member States in 2015 as an operational framework to address these challenges. The SDGs include No Poverty, Quality Education, Gender Equality, Peace, Justice and Strong Institutions, among others, as well as a meta goal on Partnerships for the Goals. Despite limitations,⁷ the SDGs form a

a <https://sdgs.un.org/goals>

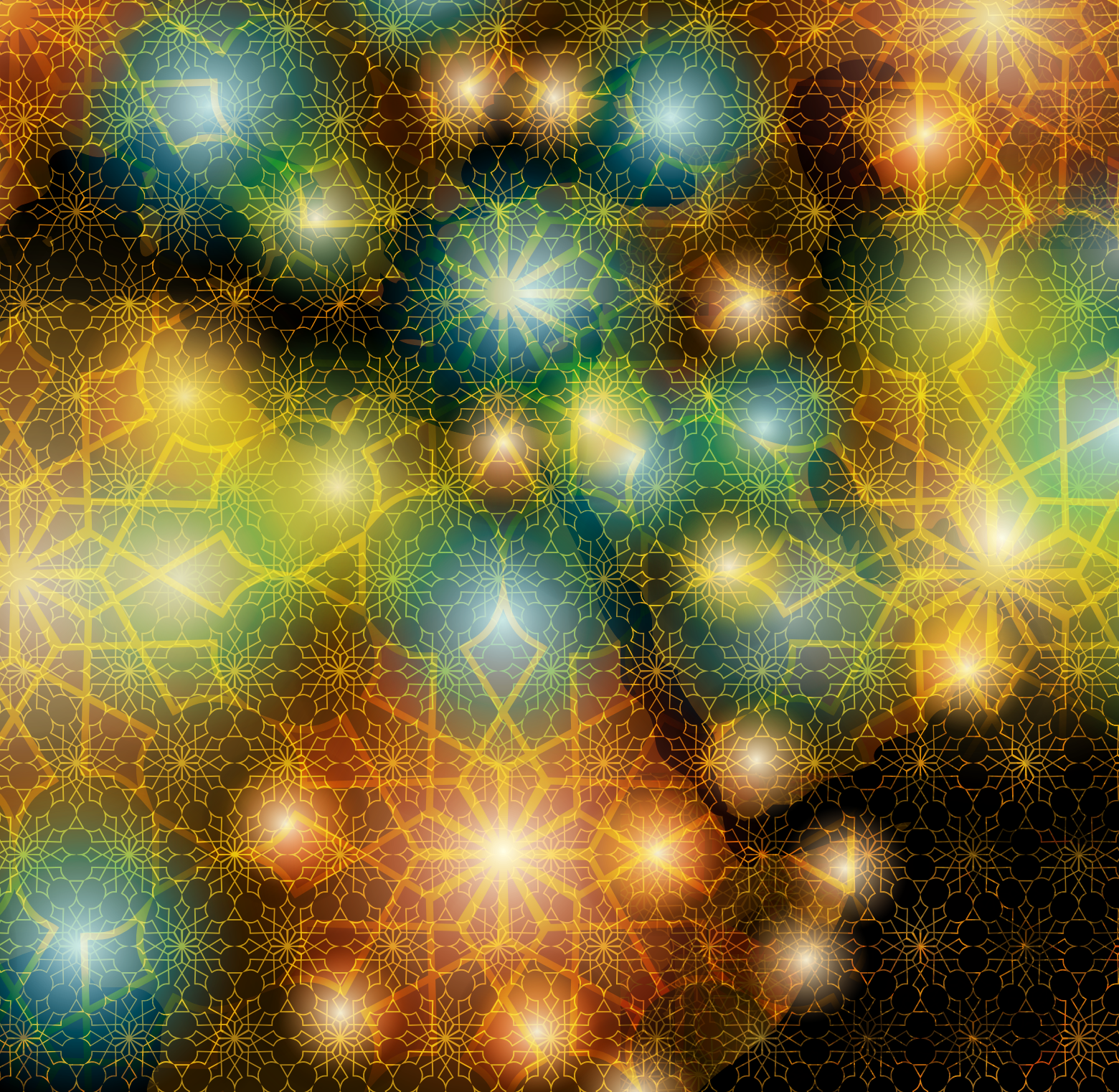
rare global consensus of all 193 UN member states on where we should collectively be heading.

Goals are meaningless without a way to track their progress. Data on the SDGs and the associated indicators^b are often outdated or unavailable, hindering progress during the Decade of Action leading up to 2030.^c Challenges around rapid access to

b <https://unstats.un.org/sdgs/indicators/indicators-list/>

c <https://unsdg.un.org/2030-agenda/decade-action>





data have also become apparent in the context of, for example, the Sudan revolution (public sentiment) or the Beirut explosion in August 2020 (infrastructure damage). The paucity of data has been highlighted during the COVID-19 pandemic, with its sudden impact on all aspects of life, most of which have yet to be quantified. Going beyond availability, accessibility, and timeliness, there is a need for more disaggregated data, such as by gender and town.

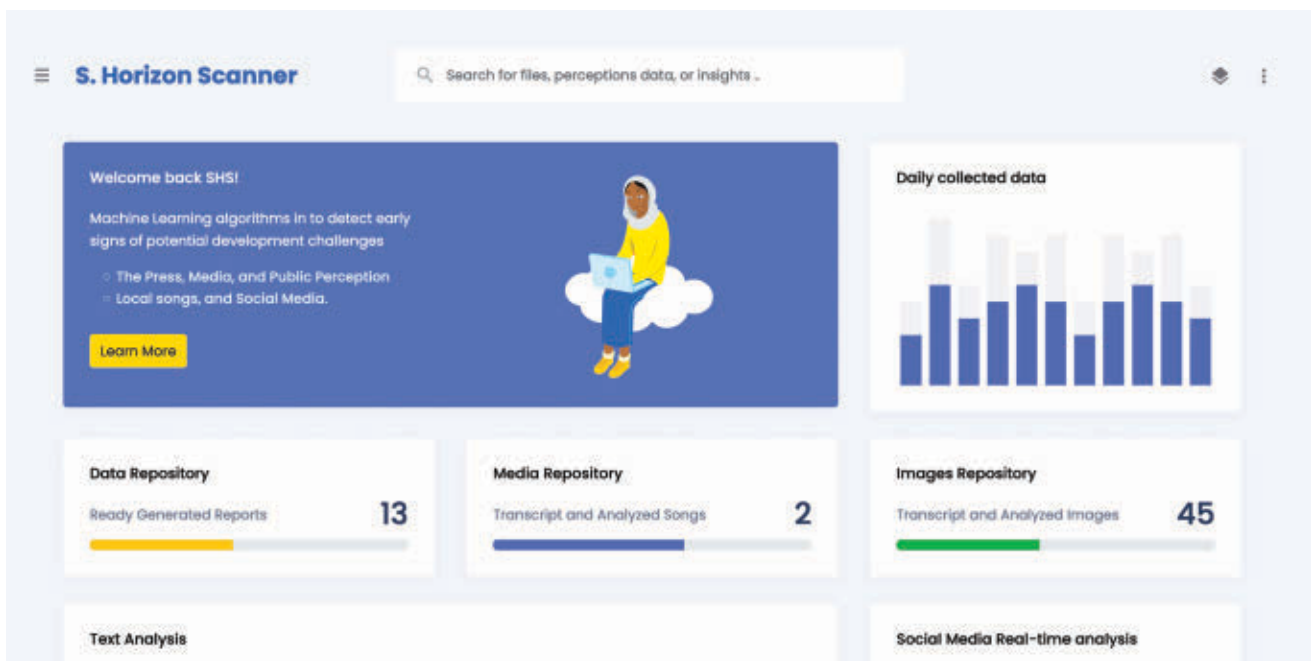
These challenges provide an op-

portunity to use non-traditional data sources as a complement to existing data and approaches, in particular through the use of artificial intelligence (AI). At the same time, a naive belief in AI as a savior risks ignoring complex, structural root causes. Furthermore, a reliance on digital traces, such as mobile phone data, risks excluding the most vulnerable—and often least connected—further aggravating inequalities. Lastly, there is a risk of taking a reductionist one-size-fits-all approach, often with a West-

ern lens and without understanding local context, in particular in the Arab region with its diverse cultures and languages.

Here, we showcase regionally developed projects that explore the use of non-traditional data sources and AI to help measure progress toward the SDGs. Some of these projects also support countries in other parts of the world, demonstrating that the Arab world is not only a consumer of, but a contributor to, world-leading innovation.

Figure 1. Sudan Horizon Scanner: A Web application that collects different types of data and media files, then applies machine learning algorithms to detect early signs of opportunities or challenges in Sudan's development.



Leveraging Behavioral and Humanitarian Data Sources to Analyze Development Challenges Faced by Syrian Refugees and Host Communities in Lebanon

Project stakeholders: UN ESCWA, Data-Pop Alliance, Qatar Computing Research Institute, Lebanese Central Administration of Statistics, Lebanese Ministry of Telecommunications, UNHCR

The need to produce timely, granular, and cost-effective estimates for the vulnerabilities faced by refugees and host communities in Lebanon remains an important priority. These estimates are normally generated through official government data and the UN High Commissioner for Refugees' register of refugees. However, families eligible for services are often not captured in the data; for example, according to the Vulnerability Assessment of Syrian Refugees (VASyR) report, only 44%¹⁶ of Syrian refugee families eligible for multipurpose cash assistance were provided with help.

In this project, UN ESCWA, in partnership with the Qatar Computing Institute (QCRI) and the Data-Pop Alliance,^d explores the potential of non-traditional data sources to gener-

ate higher quality data that can lead to more targeted service provision by hosting governments, international organizations, and NGOs, particularly at the nexus of SDG1 (No Poverty), SDG8 (Decent Work and Economic Growth), and SDG10 (Reduced Inequalities).

This project covers several data sources: call detail records (CDRs), Facebook advertising data,⁶ the Global Database of Events Language and Tone (the GDELT project),¹¹ and Twitter data. Here, we describe findings derived from CDRs.

Mobile phone metadata, such as the number and timing of outgoing calls, mobility patterns, and data consumption behavior, can be good predictors of socioeconomic status.^{2,e} In our case, CDRs from two mobile operators, Alfa and Touch, were analyzed through the Ministry of Telecommunications. Data was obtained for 12 kazas (districts), spanning the two muhafazahs (zones) Bekaa and North, for the period April-June 2016 through 2019. Data from Touch con-

sisted of the number of outgoing and incoming calls, disaggregated by age and gender; data from Alfa contained statistics about data consumption.

Overall, the spatial distribution of calls matched well with the population distribution in the CASLFS ($R^2 = 0.9$).^f Concerning spatial variation in socioeconomic, the ratio of dial-out-duration to receiving-in-duration in Touch data turned out to be a good predictor with an R^2 of 0.36 with the percentage of Syrian refugee families with debt greater than USD \$600 (using Tawasol data and VASyR ground truth⁸), and an R^2 of 0.72 with the percentage of self-declared poor in the Lebanese host community (using general Touch data and CASLFS ground truth). Details can be found in the forthcoming full report.

Accessing CDRs carries with it several questions about privacy and data governance. The collaboration between UN ESCWA and the Central Administration of Statistics was critical to ensuring the accountability necessary for the Ministry of Telecommunications to grant data access.

e See The Handbook on the Use of Mobile Phone Data for Official Statistics; <https://unstats.un.org/bigdata/events/2019/jakarta/Handbook%20on%20Mobile%20Phone%20Data%20for%20official%20statistics%20-%20Draft%20Nov%202017.pdf>

f Central Administration of Statistics Labour Force and Household Living Conditions Survey, 2018 and 2019.

g Al Tawasol was a dedicated line, hosted by Touch, for Syrians living in Lebanon.

d <https://datapopalliance.org/>

This project acts as a building block to consolidate the trust and necessary mechanisms to further explore the usefulness of this data source for the analysis of the living conditions of Syrian refugees and host communities.

Creating the Sudan Horizon Scanner for Detecting Real-Time Change

Project stakeholders: UNDP Sudan, Republic of Sudan's Ministry of Labour and Social Development, Republic of Sudan's Ministry of Trade, Republic of Sudan's Prime Minister's Office

Over the past 18 months, Sudan has “witnessed the people’s revolution and history-making transition process.”^h Along with this, Sudan has experienced a rapid change in public sentiment, a narrative that has been difficult to capture using traditional data, which were detecting neither the dynamics nor drivers of this change. As part of its sensemaking efforts, UNDP Sudan’s Accelerator Labⁱ developed the Sudan Horizon Scanner (SHS), a system to monitor a changing public narrative through real-time change detection, topic identification, sentiment classification, and summarization (see Figure 1).

The Accelerator Lab is exploring whether the system will be able to eliminate noise with minimal intervention and to explain changes in public sentiment by connecting different types of data, including socioeconomic and health data, Facebook posts, and newspaper headlines. The Accelerator Lab is also including in its analysis popular underground songs, radio shows and call-ins, and Friday prayer sermons—unusual sources of data that have captured the attention of other UNDP country offices as they look to analyze rapidly changing public sentiment in their countries.

The songs and sermons have been the most effective thick datasets for detecting signals of change. A challenge in analyzing this data is the fast rate at which vocabulary and lyrics are changing, with no existing training data for the use of these songs’

^h <https://www.sd.undp.org/content/sudan/en/home/blog/2020/reflections--my-journey-in-sudan.html>

ⁱ <https://acceleratorlabs.undp.org/>

sub-languages that are colloquially called Randook. To develop the required natural language processing (NLP) functionalities, we are therefore building our own thesaurus and training data.

Two preliminary insights derived from the Sudan Horizon Scanner include:

a. There is a strong positive correlation between higher COVID-19 infection rates and frequent water cuts in Khartoum State.

b. COVID-19 is more likely to be perceived as a conspiracy in areas with higher consumption of Kaftans, a piece of white cloth used to cover the diseased within the Muslim tradition. These areas also had higher death rates during the first peak of COVID-19.

Currently we are extending the SHS data capture and processing methods to help fill data gaps on the SDGs, particularly in the areas SDG2 (Zero Hunger), SDG6 (Clean Water and Sanitation), and SDG16 (Peace, Justice and Strong Institutions). The data will be translated into a real-time country performance tracker and monitor.

Using Social Media for Sentiment Analysis to Overcome COVID-19 Lockdown in Tunisia

Project stakeholders: UNDP Tunisia

With the onset of COVID-19, our collective reality changed irrevocably and along with it our approach to development. Values of transparency and accountability, participation and ownership should not, however, be compromised. To overcome lockdown and social distancing constraints, the United Nations Development Pro-

Mobile phone metadata, such as the number and timing of outgoing calls, mobility patterns, and data consumption behavior, can be good predictors of socioeconomic status.

Figure 2. Sentiment analysis process.



Challenges around rapid access to data provide an opportunity to use non-traditional data sources as a complement to existing data and approaches, in particular through the use of artificial intelligence.

gramme (UNDP) in Tunisia is using digital tools to inform priorities for its upcoming five-year plan through social sentiment sensing (see Figure 2).

As of January 2020, there were 6.5M Facebook users in Tunisia,^j equivalent to 55% of the population. This observation led us to study the behavior of Tunisians on social networks, particularly Facebook, to gauge trends and sentiments relating to development challenges. We collected two datasets related to Mosaïque FM,^k a private radio station in Tunisia. The first consists of 99k Arabic news headlines with their descriptions, dates, and categories. The second consists of 221k comments posted on the Mosaïque FM Facebook page, including information on titles, authors, and comments. The data are further organized by topic, such as education, politics, economy, transportation, and health.

Tunisians communicate in French, standard Arabic, dialect, Roman numerals, and emoticons. Content on social networks is characterized by orthographic heterogeneity and lack of normalization particularly in written dialects that complicate the use of NLP tools. Additionally, Arabic users often use code-switching with Latin script to communicate in Arabic, a language rich in rhetorical characteristics, vocabulary, and implicit meanings. Due to these challenges, we annotated a corpus of 22,025 training instances, including dialectal Arabic, to train a sentiment classification model.

This approach aims to guide planning and decision making through real-time analysis of public sentiment on various socioeconomic challenges and opportunities. For example, we identified strong signals about the health conditions in some regions where vulnerable populations expressed urgency in prioritizing action on SDG3 (Good Health and Well-Being). In addition, the sentiment analysis informed the development of UNDP Tunisia's upcoming program (2021–2025). Going forward, the tool will be used by UNDP managers on

a regular basis to monitor the evolution of trends and inform our work on SDGs (particularly No Poverty, Gender Equality, Reduced Inequalities, Climate Action, and Peace, Justice and Strong Institutions). UNDP Tunisia's sentiment analysis framework has been shared with other UNDP country offices across the region and around the world as an effective example of gathering inputs and generating insights in a real-time manner when face-to-face consultations of regular citizens are disrupted, in this case due to COVID-19.

Monitoring Education Insecurities

Project stakeholders: Qatar Computing Research Institute, Education Above All,^l United Nations Centre for Humanitarian Data^m

Attacks on education, such as using force against students and educators to prevent their access to education, have intensified in recent years, especially in the global South,ⁿ slowing progress on SDG4 (Quality Education). Challenges that hinder authorities' response to such incidents include missing and delayed data access. Despite the UN Security Council's monitoring and reporting mechanisms to gather timely information on violations against access to education, most of the incidents remain unreported.¹

Social networks, particularly Twitter, surface real-world events which otherwise receive limited coverage in traditional news media.¹⁷ This work seeks to capture reports of attacks on education from Twitter in Africa and the Middle East. Tweets are captured using language-specific keywords curated by domain experts. Over a period of 17 months (May 2019–September 2020), a total of 314K, 15.2M, and 161K tweets have been collected in Arabic, English, and French, respectively.

Keyword-based filtering alone is insufficient to identify pertinent reports from social media, particularly in the Arab region due to significant dialectal variations between nations. To overcome this issue, we trained

j <https://datareportal.com/reports/digital-2020-tunisia>

k <https://www.mosaiquefm.net>

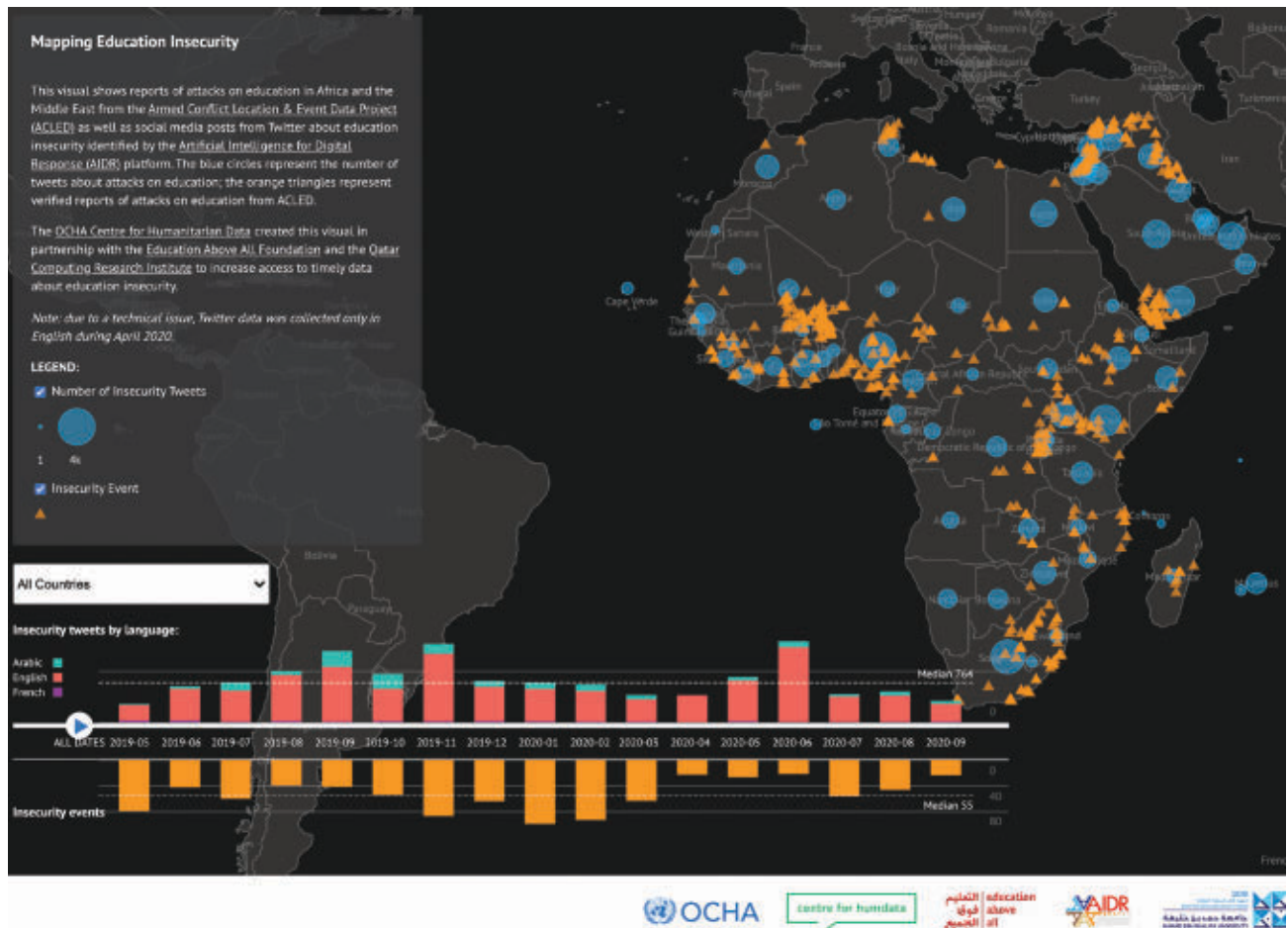
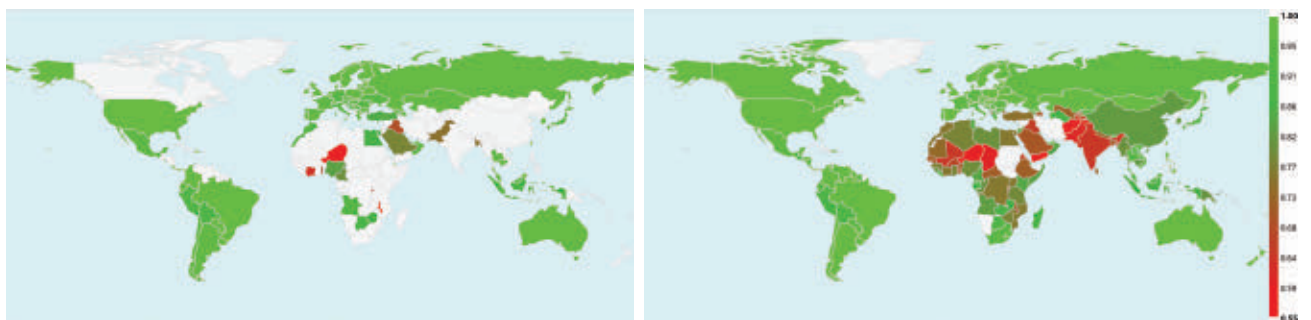
l <https://educationaboveall.org/>

m <https://centre.humdata.org/>

n <https://eua2020.protectingeducation.org/>

Figure 3. Public dashboard showing reports of attacks on education in Africa and the Middle East.

The blue circles represent the number of tweets about attacks on education; the orange triangles represent reports of attacks on education from the Armed Conflict Location & Event Data Project.¹²

**Figure 4. Female-to-male ratios of Internet users; left, taken from the most recent 2018 ITU data;⁹ right, predictions for the online model as of Nov. 11, 2020.**

three random forest binary classifiers to separate tweets “related to education insecurity” from “not related” using the Artificial Intelligence for Disaster Response (AIDR) system,⁸ which provides effective ways to collect, label, and train classifiers

using active learning. Models trained achieved AUC scores of 0.85, 0.85, and 0.79 for Arabic, English, and French, respectively.

The models reduce irrelevant reports and retain only 26K (8.4%), 5M (34.2%), and 24K (15.2%) of Arabic,

English, and French reports, respectively. We further discard tweets that are outside Africa and the Middle East, retweets, and duplicate tweets. This yields the final set of reports corresponding to 1.9K (0.62%), 11K (0.08%), and 236 (0.15%) of overall

Extreme poverty will not be eradicated through more advanced technology, and lack of data is not the reason for lack of action on climate change.

Arabic, English, and French tweets, respectively.

We deployed the system to collect and identify attacks on education reports in real-time. Pertinent geotagged reports captured by the system are being continuously shared with the stakeholders using the UN Office for the Coordination of Humanitarian Affairs’s Humanitarian Data Exchange platform^o and a public dashboard (Figure 3).^p The trained models and the system can be deployed in other Arabic, English, and French speaking countries, noting that fine-tuning may be required depending on dialectal variations in the target language.

Mapping Digital Gender Gaps Using Advertising Data and Machine Learning

Project stakeholders: Qatar Computing Research Institute, University of Oxford, Data2X,^q Sustainable Development Solutions Network

SDG5 (Gender Equality) includes a target to “enhance the use of enabling technology, in particular information and communications technology, to promote the empowerment of women.”^r The measurement for this target is incomplete, with the latest 2018 data by the International Telecommunication Union (ITU)⁹ offering statistics for female-to-male ratios (f-to-m) of Internet users for only 83

countries.^s To help fill this data gap, we launched Digital Gender Gaps, a collaboration between the Qatar Computing Research Institute and the University of Oxford, with support from Data2X.

In this project, we regularly collect data from Facebook’s Marketing API (free of charge) on the numbers of monthly active Facebook users in a given country, disaggregated by gender. We find that the f-to-m ratio of Facebook users in a given country is a good predictor of the f-to-m ratio of Internet users (adjusted R² 0.69, “online model”), better than relying on traditional offline indicators related to economic development or educational attainment (adjusted R² 0.62, “offline model”).⁴ Combining the two types of data improves the model fit (adjusted R² 0.79, “combined model”), but lowers the number of countries that predictions can be made for, as two different data sources need to be available (see Figure 4).

For the Arab region, we were expecting our model to be biased and underpredict the true f-to-m ratio as, due to cultural factors, women with Internet access might choose to refrain from Facebook more than men. We indeed observed such a gap for Oman, with an ITU reported f-to-m value of 0.93 vs. an online model

o <https://data.humdata.org/dataset/mapping-education-insecurity>
 p <https://data.humdata.org/visualization/mapping-education-insecurity/>
 q <https://data2x.org/>
 r <https://sdgs.un.org/goals/goal5>

s All female-to-male ratios considered throughout have been corrected for gender skew in the offline population. In terms of uncorrected numbers, countries such as the UAE or Qatar with a large predominantly male migrant worker population trivially have more male than female Internet users.

Wealth Index prediction performance of various models using different feature combinations of Facebook, satellite, regional indicators, and population density. All models are evaluated using R² based on 10-fold cross validation.

Model	Philippines			India		
	All	Urban	Rural	All	Urban	Rural
FB	0.595	0.446	0.451	0.483	0.235	0.393
Satellite	0.602	0.406	0.488	0.662	0.314	0.618
FB+	0.631	0.454	0.510	0.627	0.338	0.610
Satellite+	0.641	0.447	0.531	0.705	0.364	0.690
FB&Satellite+	0.655	0.464	0.546	0.708	0.368	0.695
# of clusters	1,205	437	768	28,043	8,429	19,614

prediction of 0.79 (on July 1, 2019 approximately the time point of the offline data collection), and Egypt (0.80 vs. 0.75). However, for Saudi Arabia (0.67 vs. 0.68), UAE (0.96 vs. 0.98), and Iraq (0.52 vs. 0.63) such gaps did not exist. As such our model does not seem to be systematically biased for the wider Arab region.

The project's website⁴ regularly updates its predictions based on changes in the f-to-m ratio of Facebook users. The data is also included in SDGs Today, a global hub for real-time SDG data⁵ that is used by both advocacy groups and policymakers. Going forward, we plan to extend our predictions to subnational regions.

Combining Data Sources for Mapping Poverty

Project stakeholders: Qatar Computing Research Institute, World Bank, UNICEF, Thinking Machines⁶

Non-traditional data sources such as satellite imagery^{3,10} and CDRs² have been used to map poverty at scale. We improve the state of the art by combining publicly accessible, anonymous advertising data with satellite imagery. Similar to the previous project, we use Facebook's Marketing API to obtain estimates of the proportion of Facebook users utilizing a variety of network connections (3G, 4G, Wi-Fi), and mobile operating systems (iOS, Android) in a given location. We then combine this information with other variables such as population density and features extracted from daytime satellite imagery.^{5,15}

We test our approach in the Philippines and India. As ground truth, we use the Wealth Index (WI), an asset-based measure of poverty derived from the Demographic and Health Surveys.¹⁶ We apply ridge regression to predict WI using different combinations of Facebook and satellite imagery features (see the accompanying table).

In the Philippines, with relatively high Facebook penetration (>70%), we observe that a model using only Facebook features performs on par


with a model using only satellite features. Both models improve when additional features such as population density and regional indicators are included, and the best performance is achieved when all features are combined. A geographic disaggregation shows that Facebook-only models perform better in urban locations, with satellite-only models performing better in rural locations. However, in India, Facebook features do not improve the performance of satellite-only models, neither in urban nor in rural areas.

Using Facebook advertising data for poverty estimation in countries in the Arab region with high Facebook penetration, such as Libya (>70%), appears promising. Furthermore, the advertising data affords the ability to obtain gender- or age-disaggregated poverty estimates.⁶ To encourage further evaluation of combining Facebook advertising data with other data sources for poverty mapping, this line of work has been presented at the UN World Data Forum and World Bank seminars.

Closing Thoughts

While striking a balance between case studies with a regional focus and those with a focus beyond the Arab region, all the initiatives presented here showcase regionally developed technology. Even for projects with a purely regional implementation, the lessons learned, and the knowledge obtained are disseminated throughout the global UN system, and together they offer an excellent demonstration of the opportunities that non-traditional data sources combined with AI provide for measuring and advancing the SDGs.

At the same time, these new approaches create challenges, including how to safeguard privacy and how not to exclude people without Internet connectivity, while amplifying the voices of the already-connected. More fundamentally, it is important to note that exclusion often extends beyond the data to the process of building and deploying technology. But any technology is only as good and as fair as the socio-political system it is embedded in. Put simply, extreme poverty will not be eradicated through

more advanced technology, and lack of data is not the reason for lack of action on climate change. It is now more needed than ever to broaden the group of people who build technology for the SDGs, but also who get to decide what to build, and how it will be used. 

References

1. Bennouna, C., et al. Monitoring and reporting attacks on education in the Democratic Republic of the Congo and Somalia. *Disasters* 42, 2 (2018), 314-335.
2. Blumenstock, G., Cadamuro, R. Predicting poverty and wealth from mobile phone metadata. *Science*, 350, 6264 (2015), 1073-1076.
3. Elvidge, C.D., et al. A global poverty map derived from satellite data. *Computers & Geosciences* 35(8) (2009)
4. Fatehikia, M., Kashyap, R., Weber, I. Using Facebook ad data to track the global digital gender gap. *World Development* 107 (2018), 189-209
5. Fatehikia, M., et al. The Relative Value of Facebook Advertising Data for Poverty Mapping. *International AAAI Conference on Web and Social Media* (2020)
6. Fatehikia, M., et al. Mapping Socioeconomic Indicators Using Social Media Advertising Data. *EPJ Data Science* 9:22 (2020)
7. Holden, E., Linnerud, K., and Banister, D. (2017) The Imperatives of Sustainable Development. *Sust. Dev.*, 25 (2017), 213-226.
8. Imran, M., et al. AIDR: Artificial intelligence for disaster response. *Proceedings of the 23rd International Conference on World Wide Web* (2014), 159-162
9. International Telecommunication Union. World telecommunication/ICT indicators database 2018. Data retrieved from the International Telecommunication Union website. <https://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx>.
10. Jean, N., et al.: Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (2016)
11. Leetaru, K., Schrodt, P.A. GDELT: Global Data on Events, Location and Tone. 1979-2012: ISA (2013)
12. Raleigh, C., and Dowd, C. Armed conflict location and event data project (ACLED) codebook. 2015
13. Salah, A., et al. Guide to Mobile Data Analytics in Refugee Scenarios: The 'Data for Refugees Challenge' Study. Springer (2019)
14. Savolainen, R. Information use as gap-bridging: The viewpoint of sense-making methodology. *Journal of the American Society for Information Science and Technology*, 57, 8 (2006), 1116-1125.
15. Tingzon, I., et al. Mapping poverty in the Philippines using machine learning, satellite imagery, and crowd-sourced geospatial information. *AI for Social Good, ICML* (2019)
16. UNHCR, Vulnerability Assessment of Syrian Refugees in Lebanon, 2019. <https://data2.unhcr.org/en/documents/details/73118>
17. Zhao, W. X., et al. Comparing Twitter and traditional media using topic models. *European Conference on Information Retrieval* (Apr. 2011), 338-349.

Ingmar Weber, Qatar Computing Research Institute, Qatar.

Muhammad Imran, Qatar Computing Research Institute, Qatar.

Ferda Ofli, Qatar Computing Research Institute, Qatar.

Fouad Mrad, United Nations Economic and Social Commission for Western Asia, Lebanon.

Jennifer Colville, United Nations Development Programme Regional Hub, Jordan.

Mehdi Fathallah, United Nations Development Programme, Tunisia.

Alissar Chaker, United Nations Development Programme, Tunisia.

Wigdan Seed Ahmed, United Nations Development Programme, Sudan.

Copyright held by authors/owners.
Publication rights licensed to ACM.

t <https://www.digitalgendergaps.org/data/>
u http://sdgstoday.org/8946bbc4090749c2aa1b6c1c80999bc6/page/page_20/?views=view_49
v <https://thinkingmachines.com/>
w <https://dhsprogram.com/>

BY CHRISTINA PÖPPER, MICHAEL MANIATAKOS,
AND ROBERTO DI PIETRO

Cyber Security Research in the Arab Region: A Blooming Ecosystem with Global Ambitions

IN A REGION where political tensions are recurrent, the *strive for security* is crucial. This applies equally to the cyberspace, where the need for cyber security is magnified by the level of digitization and technical penetration that the Arab region is experiencing. For instance, the Internet penetration rate^a is generally higher than 90% and, in some cases such as Kuwait, UAE, and Qatar, approaches 100%. As such, many Arab countries have recognized that the security of cyberspace is an integral part of their economic systems and a matter of national security. This awareness has been followed by policies and actions: In the International Telecommunication Union's (ITU) Global Cybersecurity Index,^b the states of Oman, KSA, Egypt, and Qatar rank among the top-20

a See <http://bit.ly/3bHrmY9>

b See <https://www.internetworldstats.com/stats.htm>

countries globally—with a considerable part of the Arab countries consistently ranking higher than many European countries. The strive for cyber security is a global as much as a local—and also Arab—endeavor, and the Arab region is gaining pace in cyber security research efforts and achievements. In this article, we will survey the main initiatives related to cyber security in the Arab region, report on the evolution of the cyber security posture, and point to possible Pan-Arab and international collaboration avenues in cyber security research.

Cyber security can be considered as specific to the Arab region as computing itself: Many of the threats, software and hardware developments, and industrial endeavors relating to cyber security are not exclusively tied to the region but are instead of a global character due to the nature of digitalization.

However, the political, economic, cultural, and financial contexts of Arab countries create a particular environment for facing attacks and addressing cyber security issues. The way the Arab world responds to cyber security challenges—in a broad but common understanding encompassing also trust and privacy—does not happen without tension or regional specificity: for instance, the protection of families and the respect for family life are an integral part of the Arab culture, while the strive for privacy protection is neither rooted nor strongly manifested in everyday digital life in Arab countries. Furthermore, while certain Arab countries are well known for their strong financial standing and politically stable systems—some being at the forefront of creating digital societies—others are suffering from war, instability, corruption, and poverty, which creates a heterogeneous and fragmented environment for threats and defenses on various scales.

As an example, the countries in the Gulf region share a strong dependency of their GDP on the oil and gas industry. For instance, the oil and gas sector



accounts for roughly 87% of Saudi budget revenues, 60% of Qatar's GDP, 40% of Kuwait's GDP, and 30% for UAE's GDP, to cite a few. Moreover, the production sites are typically concentrated in specific, narrow geographic regions, and represent a critical asset for the cited countries. For instance, on September 14, 2019, drones were used to attack the state-owned Saudi Aramco oil processing facilities at Abqaiq (Biqayq in Arabic) and Khurais in eastern Saudi Arabia, while in 2012 the Shamoon virus (aka W32.Dist-Track) was used against national oil companies including Saudi Arabia's Saudi Aramco^c and Qatar's RasGas.^d A group named "Cutting Sword of Justice" claimed responsibility for an attack on 35,000 Saudi Aramco workstations, causing the company to spend more than a week restoring

c See https://en.wikipedia.org/wiki/Saudi_Aramco
d See <https://en.wikipedia.org/wiki/RasGas>

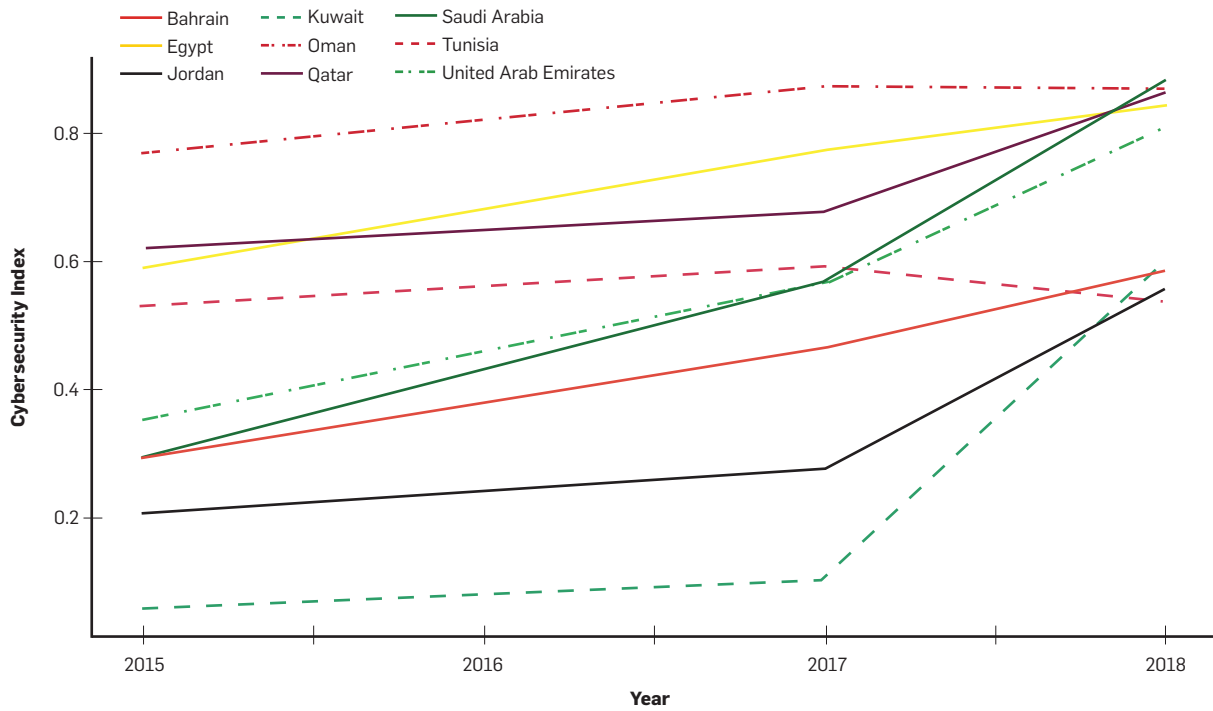
their services. Computer systems at RasGas were knocked offline by an unidentified computer virus, with some security experts attributing the damage to Shamoon. In 2017, software commonly referred to as Triton^e was the first malware to attack an industrial control system directly (not the IT infrastructure, like Shamoon did) by attacking a Saudi Arabian petrochemical plant. The cited attacks had worldwide consequences, sending up the price of oil, with further cascading effects and their increasing sophistication is alarming, pointing to state-level actors.

Consequently, awareness of the importance of cyber security raised within the national governments in the Arab region. One can observe committed endeavors toward the creation of secure digital environments within

e See [https://en.wikipedia.org/wiki/Triton_\(malware\)](https://en.wikipedia.org/wiki/Triton_(malware))

Arab countries, manifested by the development of national cyber security strategies and the establishment of national cyber security agencies—at varying levels of maturity and scope (see accompanying table). National cyber security strategies exist or are in rollout for Egypt, Jordan, Lebanon, Kuwait, Qatar and the UAE, others have occurred as drafts or are in development (Saudi Arabia, Bahrain). For other Arab countries, the recognition of cyber security as a matter requiring a national strategy is gaining momentum. The endeavors have been well directed and managed, as shown by international benchmarks. For instance, ITU's cyber security index is overall rising in many Arab countries (see accompanying figure), indicating the national strategies, capabilities, and programs in the field of cyber security are on the rise (regarding national cyber security strategies and computer emergency response teams, but also

Evolution of ITU's Cyber Security Index for Arab countries with an index > 0.5 in 2018. For comparison, the curve for Europe displays the average index of all European countries with an index > 0.5 in 2018 (averaging 40 European countries).



cybercrime legislation, awareness, and capacity building).

Despite the efforts and results described here, Arab countries continue to be popular targets for cybercriminals, partially due to their financial power and oil resources, but also due to their location in a region rife with geopolitical tensions. Since the Arab region is situated at the crossroads of vastly different cultures, it has historically been a place of major geopolitical conflict, with impact on all Arab countries. With the transition to new ways of engaging into conflict,⁷ cyber security is essential in the modern cyber-defense landscape.

The Arab Research Landscape and Initiatives in Cyber Security

The importance of cyber security research has been recognized by national governments. One testimony of Arab research efforts in cyber security are the creation of academic research centers.

In the UAE, Abu Dhabi's long-term Vision 2030 outlines to promote a sustainable, diversified, high-value-added economy and the development of a resilient infrastructure capable of supporting anticipated economic

growth. Both will depend on high-tech infrastructures and interconnected computerized systems used everywhere in the country, for example, for smart grids and intelligent transportation. Relating to the secure realization of such a vision, NYU (New York University) Abu Dhabi's Center for Cyber Security, founded in 2012, pushes frontiers with respect to academic and industrial cyber-security research, with focus on hardware security, smart city security, wireless security, critical infrastructure security, and trust/privacy. Collaborative cyber security research is conducted together with the Center for Cyber-Physical Systems at Khalifa University.

On the northeastern coast of the Arabian Peninsula, the State of Qatar issued the Qatar National Vision 2030 that recognized cyber security as a strategic pillar. In this respect, the most recent effort is the creation of the Qatar National Agency for Cybersecurity that would provide a unified center for coping with cyber security threats. This last effort adds to the Qatar Computing Research Institute (QCRI) at Hamad bin Khalifa University (HBKU) that sports a Cyber

Security department, featuring research activities aimed at supporting governmental needs. Further, the very same HBKU has supported, within its College of Science and Engineering located just at the outskirts of Doha, the cited cyber security pillar creating the Cybersecurity Research and Innovation Lab that conducts research on a broad range of topics, with a focus on critical infrastructures protections (protection of avionics communications, maritime communication, drones and satellite security, and so forth). Further research efforts in cyber security are carried out at Qatar University and Carnegie Mellon University Qatar, mainly in the field of secure computing.

In Saudi Arabia, King Abdullah University of Science and Technology (KAUST) is pursuing a university-wide cyber-security initiative and is creating faculty positions at all levels in fields related to cyber security. The role played by KAUST in cyber security is magnified by the recent memorandum of understanding^f signed with

^f See <https://www.kaust.edu.sa/en/news/advancing-cybersecurity>

Saudi Arabia's National Cybersecurity Authority aimed at strengthening the cyber security capabilities of the Kingdom's workforce, focusing on developing educational programs in cyber security and providing consultations on curricula and training courses.

These exemplary and other initiatives bear fruit in connecting international cyber security researchers to the Arab region and in hosting existing and establishing new cyber security activities. In Feb 2020, the Global Cybersecurity Forum^g took place as an international two-day event in Riyadh, Saudi Arabia, concluding with the formulation of the Riyadh Declaration for Cybersecurity,^h being a commitment to cyber security objectives. November 2020 marked the 6th anniversary of the Cyber Security Awareness Week (CSAW) MENAⁱ region, organized by NYU Abu Dhabi. In 2019, HITB⁺ (Hack in the Box) CyberWeek^j took place for the first time, being the largest HITB cyber security event in the Middle East, followed by a virtual edition in 2020. CyberTalents, an Egypt- and UAE-based company has been organizing an Arab and Africa Cybersecurity CTF (Capture the Flag) competition yearly

- g See <https://www.globalcybersecurityforum.com>
 h See <https://www.globalcybersecurityforum.com/declaration>
 i See <https://www.csaw.io/mena>
 j See <https://cyberweek.ae>

since 2017, currently covering ten Arab countries: Saudi Arabia, Oman, Sudan, Kuwait, Algeria, Morocco, Lebanon, Jordan, Tunisia, and Egypt. In 2019, Qatar's HBKU organized the first Qatar International Cybersecurity Contest, aiming at an interdisciplinary cooperation in cyber security—including social sciences, law, health sciences, Islamic studies and CS students—catalyzing more than 180 international participants. In terms of academic conferences, AsiaCCS 2017,^k ACM's Asia Conference on Computer and Communications Security, took place for the first time in the Arab region at NYU Abu Dhabi, hosting more than 200 international computer and cyber security researchers.

The liveliness of the cyber security ecosystem, and its global reach, is also testified by business success stories. For instance, DarkMatter, a UAE-based company that offers a complete portfolio of cyber security solutions underpins its work by industry-leading intelligence, research and development, resulting in the creation of the Technology Innovation Institute.^l In Qatar, the Qatar Science and Technology Park is home to a vibrant start-up ecosystem

- k See <https://dl.acm.org/doi/proceedings/10.1145/3052973>
 l See <https://www.tii.ae>

Despite the efforts and results described here, Arab countries continue to be popular targets for cybercriminals, partially due to their financial power and oil resources, but also due to their location in a region rife with geopolitical tensions.

Overview of National Cyber Security Strategies (NCSS) and National Cyber Security Agencies/Bodies in selected Arab States. (See online appendix for notes and references <https://dl.acm.org/doi/10.1145/3447741>)

State	National Strategy			National Agency	
	NCSS available	Year of Creation	Current Coverage	Body available (since)	Name of Body
Bahrain	○ ^{1appx}	—	—	in development ^{1appx}	(MoI, iGA)
Egypt	● ^{7appx}	2017 (upd. in 2018)	2017–2021	✓ (2014)	ESCC ^{3appx}
Jordan	● ^{8appx}	2012	2018–2023	in development ^{8appx}	(MoICT)
Kuwait	● ^{9appx}	2017	2017–2020	in development ^{9appx}	NCSC
Lebanon	● ^{6appx}	2019	2019–2022	in development ^{6appx}	NCISA
Oman	◐ ^{4appx}	2017	—	✓ (2010)	OCERT ^{11appx}
Qatar	● ^{12appx}	2014	2014–2018	✓ (2020/21) ^{13appx}	NACS; Q-CERT(2005)
Saudi Arabia	◐ ^{2appx}	2013*	2013–2024	✓ (2017)	NCA ^{14appx}
Tunisia	◐ ^{15appx}	2018	2020–2025	✓ (2004)	ANSI ^{16appx}
UAE	● ^{10appx}	2019	2019–2022	in development ^{10appx}	(TRA); aeCERT (2008)

The liveliness of the cyber security ecosystem, and its global reach, is also testified by business success stories.

that enjoys generous funding and excellent logistics, aimed at incubating high tech start-ups, including the next stars of cyber security. In Egypt, the world's largest FinTech accelerator (Startupbootcamp) has put its root in Cairo, to foster innovation in financial inclusion and the general startup ecosystem, while local cyber security start-ups are growing up and, in a few promising cases, are acquired by international companies to boost their growth. A similar trend is experienced in the KSA. For instance, in 2019, Saudi Arabia's startup ecosystem saw an investment of \$67m, registering a 35% increase compared to a year before. The Kingdom also witnessed an increase in government initiatives, accelerator programs, and the total number of investors as well. All in all, the Arab region cyber security start-up ecosystem shows a clear positive trend, and it is poised to blooming.

Research Highlights and Awards

Among the many facets of cyber security, notable research environments and results relate to security for smart cities and critical infrastructures, maritime and aerial transportation security, hardware security, communication and Internet security, misinformation and fake news, as well as the use of AI for cyber security:

► *Hardware security:* NYUAD has created the world's first chip considered unhackable. Researchers of NYUAD's Design for Excellence (Dfx) lab developed a 'logic-locked' security chip to protect devices from the surge in cyber-attacks. Secured by a secret key, it permits only authorized users to utilize the device and it is also resistant to reverse engineering. The group's research was presented at the ACM Conference on Computer and Communications Security,¹⁷ one of the leading cyber security conferences in the world.

► *Smart city and critical infrastructure security:* NYUAD's Center for Cyber Security hosts a smart city testbed as an Internet of Things (IoT) platform with a collection of interoperable processes, simulation models, hardware devices, and appropriate network protocols. With application to smart grid, intelligent transportation, chemical plants, smart houses/buildings, and desalination plants, it is being used

to produce a series of offensive and defensive results in industrial control system security: It was used to identify power grid vulnerabilities^m (CVE-2017-7905, presented at BlackHat USA 2017), investigate GPS spoofing attacks possibly leading to time synchronization problems, demonstrate attacks on firmware modification of power grid devices and detect malicious modifications in the firmware of embedded systems.⁸ HBKU's College of Science and Engineering is home to the Cybersecurity Research and Innovation Lab (CRI-Lab)ⁿ that also performs research on the protection of critical infrastructures. The lab has strategic plans, advanced testbeds, and top-notch researchers to address IoT security¹³ and the application of AI techniques to relevant cyber security research problems.¹²

► *Aerial and maritime security:* HBKU's Cybersecurity Research and Innovation Lab also addresses maritime cyber security,³ avionics security,¹¹ and satellite security.⁹ Researchers from NYUAD's Cyber Security & Privacy Lab have developed DeepSIM, a technique to detect GPS spoofing attacks on UAV's using camera input and satellite imagery matching¹⁶ and MAVPro, an approach for ADS-B message verification for air traffic security that increases the geographic coverage of multilateration-based verification.⁵

► *Communication security and applied cryptography:* In an international team, scientists from Saudi-Arabia's KAUST published a paper in Nature Communications⁶ to demonstrate perfect secrecy cryptography in classical optical channels^o using chaos theory and the second law of thermodynamics, a possible response to the emergence of quantum computers and the risk this carries for classical cryptographic approaches. Their approach relies on correlated mixing of chaotic waves in an irreversible time-varying silicon encryption chip. In other international collaborations with first-tier publication results, NYUAD researchers analyzed the Dragonfly handshake of WPA3,¹⁵ the recent Wi-Fi security

m See <https://www.reuters.com/article/us-cyber-generalelectric-power-idUSKBN17S23Y>

n See <https://cri-lab.net>

o See <http://bit.ly/39zr4Qt>

standard that is replacing WPA2, and introduced a new paradigm for side-channel attacks over remote connections by exploiting concurrency to leak secrets.¹⁴ Qatar University and Carnegie Mellon University in Qatar are also addressing special topics in cyber security, such as the challenging endeavor of producing a viable sound prototype of a garbled computer—an appealing alternative to homomorphic encryption.¹⁰ HBKU-CSE's research in this domain addresses privacy² and cloud computing.⁴

► **Internet security:** With its Security Department, HBKU's research institute QCRI sports a number of initiatives in the cyber security field. It actively investigates relevant problems with worldwide impact, such as MADA:^p A system for identifying malicious domains using a real-life 'guilt by association' principle. It detects malicious domains by analyzing the movements and previous associations of a domain address—it can analyze 50 million domains in six minutes.

► **Misinformation and fake news:** Another branch of QCRI leads the Tanbih mega-project,^q developed in collaboration with MIT. The project aims to build a news aggregator that limits the effect of fake news, propaganda and media bias—a key tool for the Arab region given the high Internet penetration as described at the beginning of this article, combined with the fact that the vast majority among the young Arab generations consume online news.

► **AI for cyber security:** QCRI has also helped define the AI strategy for Qatar,^r a cornerstone element to address current and future challenges of cyber security.

Cyber security research is also conducted in Lebanon, where the research group led by Ali Chehab at the American University of Beirut is addressing topics from SDN security, to physical-layer security, and network coding. At the other end of the spectrum is Egypt, where cyber security is a key topic for the industrial sector, with collaborations in place with top indus-


trial players like Siemens and Valeo.

Finally, research highlights from the Arab region include recent awards relating to cyber security. In 2020, a Google Ph.D. Fellowship in Privacy and Security went to Yunusa Simpa Abdulsalm from Mohammed VI Polytechnic University in Morocco, who is working on securing the electronic health system. With this selection, the Google Ph.D. Fellowship was awarded for the first time ever to the Arab Region on a cyber security topic. In 2020, Naif Saleh Almakhdhub, Assistant Professor of Engineering at KAUST in Saudi Arabia, received the CSAW MENA award for his NDSS 2020 paper on a compiler-based mitigation to prevent control-flow hijacking attacks for securing embedded systems.¹ In 2019, Amer Al Jaberi from the UAE received the MIT Technology Review MENA Region Innovator Under 35 distinction for a document and passport reader being used by all immigration portals in the UAE, verifying the veracity of the documents while safeguarding the various ports of entry from fraud. The young academics and engineers are raring to go.

Our Vision for the Future of Cyber Security in the Arab Region

The cyber security landscape for the Arab region presents unique challenges and opportunities. The importance of cyber security is well represented in the national policies, funding is provided to adequately address the challenges of choice, and a vibrant research ecosystem in cyber security is gaining more and more momentum. However, the efforts for creating secure digital environments and establishing research excellence in cyber security in Arab countries are still often reactive and fragmented. The high level of dependency on few key critical infrastructures (namely the oil and gas sector) and the level of threats that the countries in the region are facing, call for further collaborative actions to devise solutions idiosyncratic to this unique context.

In particular, the need for further coordination, sharing of best practices and experiences, boosting research excellence, and the creation of specific transnational research

projects aiming at solving specific issues shared by countries in the region would be highly beneficial for the whole region and will reinforce the Arab region experience in cyber security as a model for the rest of the world. This will take time but invoke irreversible benefits for the region. Finally, to unleash the Arab region potential, we emphasize the importance of an excellent cyber security training and education infrastructure, including academic cyber security programs and specializations. 

References

- Almakhdhub, N.S. et al. RAI: Securing embedded systems with return address integrity. NDSS 2020.
- Bentafat, E., Rathore, M. and Bakiras, S. A practical system for privacy-preserving video surveillance. ACSAC 2020.
- Caprolu, M. et al. Vessels cybersecurity: Issues, challenges, and the road ahead. *IEEE Commun.* 58, 6 (2020), 90–96.
- Chkribene, A. and Erbad, R. Hamila: A combined decision for secure cloud computing based on machine learning and past information. *IEEE WCNC* 2019.
- Darabseh, A., Alkhzaimi, H., and Pöpper, C. MAVPro: ADS-B message verification for aviation security with minimal numbers of on-ground sensors. *ACM WiSec* 2020, 53–64.
- Di Falco, A. et al. Perfect secrecy cryptography via mixing of chaotic waves in irreversible time-varying silicon chips. *Nature Communications* 10, 5827 (2019). <https://doi.org/10.1038/s41467-019-13740-y>
- Di Pietro, R. et al. *New Dimensions of Information Warfare*. Springer International Publishing, 1st Ed. (2021); <https://doi.org/10.1007/978-3-030-60618-3>
- Keliris, A. and Maniatakos, M. ICSREF: A Framework for Automated Reverse Engineering of Industrial Control Systems Binaries. NDSS 2019.
- Oligeri, G., Sciancalepore, S., and Di Pietro, R. GNSS spoofing detection via opportunistic IRIDIUM signals. *ACM WiSec* 2020, 42–52.
- Rachid, M.H., Riley, R. and Malluhi, Q.M. Enclave-based oblivious RAM using Intel's SGX. *Comput. Secur.* 91, 101711 (2020).
- Sciancalepore, S. and Di Pietro, R. SOS: Standard-compliant and packet loss tolerant security framework for ADS-B communications. *IEEE Transactions on Dependable and Secure Computing* (2019); <https://doi.org/10.1109/TDSC.2019.2934446>
- Sciancalepore, S. PiNCh: An effective, efficient, and robust solution to drone detection via network traffic analysis. *Comput. Networks* 168 (2020).
- Tedeschi, P. et al. LiKe: Lightweight Certificateless Key Agreement for Secure IoT Communications. *IEEE Internet Things J.* 7, 1 (2020), 621–638.
- van Goethem, T. et al. Timeless timing attacks: Exploiting concurrency to leak secrets over remote connections. In *Proceedings of the USENIX Security Symposium 2020*: (2020), 1985–2002.
- Vanhoef, M. and Ronen, E. Dragonblood: Analyzing the Dragonfly handshake of WPA3 and EAP-pwd. In *Proceedings of the IEEE Symposium on Security and Privacy 2020*, 517–533.
- Xue, N. et al. DeepSIM: GPS spoofing detection on UAVs using satellite imagery matching. ACSAC 2020.
- Yasin, M. et al. Provably-secure logic locking: From theory to practice. *ACM CCS* 2017, 1601–1618.

Christina Pöpper, NYU Abu Dhabi, UAE.

Michail Maniatakos, NYU Abu Dhabi, UAE.

Roberto Di Pietro, HBKU-CSE, Doha-Qatar.

The information and views in this article are those of the authors and do not necessarily reflect the official opinion of their institutions.

© 2021 ACM 0001-0782/21/4

p See <https://www.hbku.edu.qa/en/research-groups/cyber-security>

q See <http://tanbih.qcri.org>

r See https://qcai.qcri.org/wp-content/uploads/2019/10/QCAI_MOTC_AI_Strategy_English_FINAL.pdf

BY SLIM ABDENNADHER, SHERIF G. ALY,
JOE TEKLI, AND KARIMA ECHIHABI

Unleashing Early Maturity Academic Innovations

THE ARAB REGION consists of many teaching-intensive universities that are intrinsically committed to holistic educational excellence. According to a recent UNESCO report,⁵ the higher education sector in the Arab region is undergoing a need for massive expansion given exponentially growing populations, record-breaking youth cohorts, coupled with a strong recognition of the economic and social value of higher education. Such an enormous need for growth poses a significant challenge for publicly funded universities yet offers many opportunities for private universities to meet the ever-increasing demands of advanced education.² As is the case with many similar universities worldwide, not being dedicated research institutions often results in limited availability of research funds, resources, and hence innovation throughput. The examples given in this paper are those of universities in the region that were initially focused on consolidating their teaching,

except for one which started first as research-intensive. However, it was not long before a shift in policy included research excellence in undergraduate education by harnessing the most valuable resource of any university: the aspiring students themselves.

While the different universities followed seemingly different approaches to tap into undergraduate student potential, most successful models follow the same broad guidelines. The heart of stimulating high-quality undergraduate research innovation in computing lies in enabling full potential through early maturity, stimulation of discovery, exposure to international collaborations and projects while providing students with the needed freedom to grow and innovate. We take a shot at explaining how this is happening through four prominent universities from diverse areas across the Arab region.

The Enabling Ecosystem

Favorable conditions in the regional educational and societal system pave the way for successful undergraduate student transformations towards domain innovation, creativity, and excellence. For example, computing programs in the Arab region often attract the best students emerging from pre-university education. In most cases, the students are already enabled with a solid mathematical and scientific foundation by virtue of the anatomy of high school education and the accompanying conditions to qualify for university computing programs. On another front, computing education is trending in the region with a reputation for high market demand, a certain future, and high pay. Also, most prominent private universities in the region headhunt students by offering full and partial scholarships to top-ranking high school students. While STEM and computing specializations suffer from biased gender ratios in many regions, such as the United States and Europe, this type of education in the region is relatively gender-



The campus of the American University in Cairo, Egypt.

balanced compared to the rest of the world.³ This balance makes the skill and interest pool of the students also relatively balanced and diverse. On the other hand, limited research funding, a limited number of full-time graduate students, and a need to meet stringent requirements for academic rank progression led many university researchers to rely on undergraduate students early on. This is achieved by providing them with mentorship, incentivizing them to take on more responsibilities and tasks than usually accustomed for students at this level. Finally, a powerful desire from all involved parties for international collaboration, participation in pre-university computing competitions, and extracurricular work is deep-rooted amongst this type of pre-university student population.

The American University in Cairo

(Egypt), the German University in Cairo (Egypt), the Lebanese American University (Lebanon), and Mohammed VI Polytechnic University (Morocco) are four private universities from diverse regions of the Arab world. They are relatively small in size compared to other regional counterparts. All of them started as teaching-intensive, except for the latter, which started as a research-intensive university before incorporating undergraduate programs. In the rest of the paper, we present the four different case studies of these universities and how they realized the early maturity approach for research and innovation.

The AUC Case Study

Since 1919, the American University in Cairo (AUC) has been a beacon of academic excellence in the region. The

university was founded by Americans devoted to education and service in the Middle East, is strongly committed to liberal arts education, and strongly fosters critical thinking and creativity.^a In 1988, the Computer Science Department was founded and later renamed to Computer Science and Engineering in 2008. The two programs were designed based on the highest academic standards of counterparts at North American universities and by the professional affiliations of its alumni. Many of the program alumni are working in highly impacting leadership and technical roles at Facebook, Dell, Google, Microsoft, MongoDB, Fujitsu, SAP, Amazon, Haliburton, Cisco, General Electric, UiPath, GitHub, NASA JPL, and more. Many have also either

^a <http://www.aucegypt.edu>

completed their education or have occupied academic positions at many leading universities around the world including Harvard, Carnegie Mellon University, Stanford, Boston University, Cambridge University, University of Alberta, Texas A&M, Northeastern, University of Waterloo, Columbia, MIT, and more. Others have initiated startups including the renowned Affective, Agolo, and Voicea. In 2018, the computing programs went through further radical changes to give more competitive edge to its graduates that included:

Early maturity. Course sequences were re-engineered to provide students at least one semester worth of earlier than usual maturity of technical content. Specifically, the first two fundamentals of computing courses that students take during the first two semesters in either Computer Science or Engineering were notably redesigned to accommodate more knowledge units. This included an in-depth study of problem-solving using C++ for a systems flavor, a notable coverage of data structures in both courses, more real life problems, and two formalized lab hours with both courses to offer a solid, mentored hands-on experience. Unnecessary prerequisites were removed to open opportunities for earlier than usual branching in the program of study, including an ability to take algorithm analysis and design (a key enabler) during the third semester as opposed to the fifth semester of study, as well as operating systems soon after. The idea of early maturity was to allow students an opportunity to learn and achieve earlier before full-time engagement into the job market.

The Exploration Studio. Generous funds were made available to allow learners to experiment with computational technologies of their choice. At the end of exploration, students were required to “share and enjoy” their know-how. This provided students an opportunity to experiment with novel tools, share lessons learned, and also to help build an infrastructure for use by others. Within two years of inception, 16 proposals were submitted by groups of students of about three to five students each, some of which were required to be interdisciplinary. The funded proposals were in many areas

including but not limited to autonomous vehicles, cybersecurity, nanotechnology computing, embedded systems, and IoT.

Capstone empowerment. Besides rigorous technical requirements, capstone (thesis) projects must satisfy one or more empowerment outcomes: (a) produce a publishable innovation; (b) take a clear path toward productization; (c) serve a community purpose. Students were provided with sessions to boost their research and entrepreneurial capabilities. In Fall 2020, over half of the projects had a research element to them, so the standard of the project technicalities was notably boosted compared to previous semesters, and students were very engaged in undergraduate research activities. Projects in Fall 2020 alone included innovations in precision agriculture, healthcare for the elderly, fake news detection, human activity recognition, brain-computer interfacing, coronavirus diagnosis, visual speech recognition, and social augmented reality.

Entrepreneurial immersion. The Department of Computer Science and Engineering partnered with the university’s Business School, and the first-ever satellite location of the university venture lab was created within the premise of the department. The idea was to bring the startup culture as close and as early as possible to the undergraduate student community so that students can observe and interact with counterparts in startups, and for the startups to export their technical challenges to the booming undergraduate population.

SPRITE branding. The programs were rightly branded as being Student Oriented, Popular, Research Intensive, Industrially Aligned, Technologically current, and Experiential. This was made visible in program activities so that the constituents of the computing programs were well aware of the vision.

Industrial training. The bar was raised for expectations related to industrial training so that students can benefit from an eight-week-long internship. Undergraduate students were also put in contact with alumni of the programs from prior years as a means of facilitating internships and employment.

Faculty. New vibrant, industrially experienced faculty were recruited to complement the existing body of faculty, some of whom were graduates from Ivy League universities with strong interdisciplinary backgrounds, and others with notable Silicon Valley industrial experience.

The GUC Case Study

The German University in Cairo (GUC) is an independent, private, non-profit Egyptian institution, led by a consortium of Egyptians and Germans focused on multidisciplinary cooperation between the two countries. It is the first integrated university outside Germany offering B.Sc., M.Sc., and Ph.D. degrees in 71 study programs. In its startup phase in 2003, the primary focus of the GUC was providing excellent teaching quality. After consolidating the teaching part, the policy shifted to follow the Humboldtian model^{1,4} with a dual mission of research and teaching with unbiased and independent current research trends guiding curricula. The university continually expands in multiple senses in response to the increased teaching demand. All university members, starting from the top faculty to the youngest students, are involved in this institutional framework of research and teaching unity. The GUC acknowledges the need for early involvement of young promising talents in the whole research and education lifecycle, where students are not only provided with technical and professional skills but are also allowed to seek growth and enabled to carve their education and professional path. Several diverse measures realize the following:

Research-oriented education. The curriculum consists of research- and industry-conscious courses, for example, the programming labs of each semester and the software engineering course. Most advanced courses (including electives and pre-masters) have research projects integrated into them with requirements that are adjusted according to current research trends and market needs. The course content is also flexible, allowing the integration of needed concepts and technologies.

Early publication motivation.

Undergraduate students exploring research projects on their own or involved in bigger ones are mentored to publish their work in international conference proceedings. Bachelor thesis students are also encouraged to target publications based on their projects. Over the past five years, 50% of the GUC ICT publications resulted from undergraduate research.

Research environment exposure.

Through various internal and external funds, students are exposed to numerous opportunities for participating in research and technical environments (not only through conventional internships). Over the past three years alone, the Faculty of Media Engineering and Technology hosted more than 25 international conferences, hackathons, workshops, and summer schools for students, promoting state-of-the-art technologies such as brain-computer interaction,^b virtual reality, research best practices, and programming concepts. The GUC hosted the 17th International Conference on Mobile and Ubiquitous Multimedia (MUM'18),^c as well as a number of workshops at international conferences including MUM'17, MUM'18, KI'19, and PAAMS'20.^d

Involvement in cooperative research. As soon as students enter the university, they are offered the chance to join multiple research groups that enable them to join international research trips. This exposes them to international partners and various opportunities for enhancing studies through extended research trips and internships at partner institutions. The research groups' focuses include virtual and augmented reality, technology in education, self-driving cars, natural language processing, and character computing which take place in five dedicated labs (3D Lab, iLab, Cube, IoT Lab, and Self-driving Lab).

Startups and entrepreneurship.

Many students who take the lead in shaping their research projects aspire to turn them into a startup. They find guidance and mentoring opportunities enabling them to seek such independence. Start-ups resulting

from the GUC include Robusta,^e Null Dies,^f thndr,^g and DREIDEV.^h Many undergraduates work part-time in the industry or as freelancers.

Bachelor thesis abroad. Students are the bridge for joint research with German institutions and this international outreach. The bachelor thesis exchange program has been ongoing for the past 13 years, with a yearly average of 15% of students performing their bachelor projects at partner institutions in Germany and Europe. The same applies to funding opportunities by the German Academic Exchange Service for bilateral postgraduate studies abroad, thus increasing students' interests in an academic career early on.

Academic freedom. Independence is encouraged by enabling interested students to pursue their research directions and extracurricular courses, thus helping them become autonomous by developing their own capabilities and paths. This includes suggestions for seminar topics, elective and research topics, for example, courses related to knowledge required for self-driving cars, NLP problems in code-switched Arabic-English, and affective and character computing.

The LAU Case of Student Centeredness

The Lebanese American University (LAU) was founded as the American School for Girls in 1835, and the institution evolved into a full-fledged university in 1973, and later became the Lebanese American University in 1995. It currently boasts three campuses in Byblos and Beirut housing seven academic schools and more than 300 faculty and 8,000 students. Founded in 1995, the School of Engineering (SOE) is LAU's crown jewel, leading the university in academics, innovation, administration, and research quality. The SOE has established a model for student-centeredness, revolving around six main pillars: empowering students, motivating students, promoting student collaborations, promoting



STEM and computing specializations in education are relatively gender-balanced in the Arab region compared to the rest of the world.



b <https://www.br41n.io/>

c <http://www.mum-conf.org/2018/>


d <https://www.paams.net/workshops/c2>

e <https://robustastudio.com/>


f <http://nulldies.com/>

g <https://thndr.app/>

h <https://www.linkedin.com/company/drei-dev/>



The idea of early maturity was to allow students an opportunity to learn and achieve earlier before full-time engagement into the job market.



undergraduate research, showcasing student achievements, and securing professional opportunities.

Empowering students through academic and professional clubs, encouraging them to highlight their scientific and social talents through a dozen vibrant student clubs. SOE currently houses 10 vibrant student clubs, grouping more than 800 students from all engineering majors: ASME, ASCE, AI, Emergent Technologies, Engineers Without Borders, IEEE, IISE, Google DSC, Robotics, and SPE. Since 2018, the clubs have organized around 40 activities (four per club) every year, including workshops, invited speakers, excursions, and social events, attracting large crowds of participants from inside and outside campus.

Motivating students. High-impact hackathons and competitions are organized frequently with industrial partners, involving a large number of students in intensive competition type environments, while connecting with international partners for exposure and experience sharing. In 2018–2019, the SOE organized one international and four national completion events,ⁱ involving more than 2,500 student participants: *First Lego League (FLL) 2019 international robotics championship*, organized for the first time in the MENA region (1,000 participants); *National Education Robotics Day (NERD) Open 2018* (400 participants); *NERD National 2019* (700 participants); *BMW Group Beirut Hackathon 2019* in collaboration with Nvidia and Oracle (72 participants); and the *Popsicle Stick Bridge Competition 2019* (250 participants).

Promoting interdisciplinary student collaborations. SOE and the School of Medicine have worked together to create a new interdisciplinary collaboration. For example, since 2018 medical students can initiate collaborative projects with engineering students in *Computational Health Informatics*. Other initiatives include a series of *E-mobility* projects, gathering students to work on projects related to electric transportation systems.

i FLL 2019: <https://bit.ly/3m3Lu8w>, NERD Open 2018: <https://bit.ly/37RjpHO>, BMW Beirut Hackathon 2019: <https://bit.ly/3a5Abdw>

SOE organized the first road tests of hybrid and electric vehicles in Lebanonⁱ in January 2019 to assess their performance under local conditions. Also, the school annually organizes the LAU Engineering Week including invited talks, workshops, activities, and an annual gala dinner to celebrate engineering.

Promoting undergraduate research. SOE is attempting to harness the energy of undergraduate students by launching undergraduate research courses in every engineering program and a research methods course in the honor's program. While the former allows motivated students to get a foretaste of research activities, the latter allows them to study the theoretical foundations of the research methodology and apply it in their senior projects. Since 2018, more than 35 undergraduate students have participated in research projects, most of them publishing their papers in international conferences or journals.

Showcasing student achievements. SOE has established many annual events aiming at showcasing student achievements. One such event is the *SOE Pioneers Day*, which invites distinguished SOE students and alumni to present their projects and experiences. Another event is *SOE Projects Day*^j allowing SOE students to present their capstone projects in the presence of representatives from major industry partners (including Murex and CME Offshore, among others). SOE holds one of the highest student employability rates in LAU, with almost all students securing jobs within six months following their graduation.

Securing professional opportunities. Through SOE's Career and Placement Office and LAU Alumni Relations Office, the university helps its students secure the best internship and job opportunities in the market. One of the strongest internship programs has been established with the BMW Group automotive giant, where more than 60 students since 2018 have completed internships at the company's Logistics Robotics department in Munich.^j SOE has also built strong ties with many multinational companies, including

j <https://news.lau.edu.lb/2019/inside-the-bmw-internship.php>

FEV France, Murex, FOO, InMind.ai, Mitsulift, among many others.

The UM6P Case Study

Morocco boasts a large number of teaching-centric institutions delivering undergraduate and graduate computer science degrees. In 2017, Mohammed VI Polytechnic University (UM6P) was established as the first research-intensive university in the country, with innovation integrated into its core. Below, we identify some key factors that helped boost innovation at the undergraduate level.

Graduate education foundations first. UM6P schools brought graduate programs to a level of maturity first before starting undergraduate programs. This helped each school embed a culture of research and innovation, which seamlessly propagated to undergraduate programs. For example, the School of Computer Science (UM6P-CS) launched a fully funded four-year Ph.D. program in 2018 and will offer a subsequent undergraduate degree in computer science by fall 2021. There are currently 36 fully funded Ph.D. students enrolled at UM6P-CS, one of whom holds a Google Ph.D. Fellowship. Research efforts by UM6P-CS students and faculty have led to publications in top venues.

Attracting and retaining top talent. UM6P-CS recruits top students by offering generous scholarships and a world-class campus experience. Moreover, it attracts faculty with notable international research experience by providing them ample startup funds, the financial autonomy to build research lab(s), and a reduced teaching workload. A large network of national and international partnerships allows students and faculty members to collaborate on large-scale projects with real potential to drive societal change. UM6P-CS collaborates with prestigious institutions worldwide on a number of research projects, including *Byzantine-Resilient Coordinate Descent* and *Content-Agnostic Fake News Detection* (EPFL), *Scalable Machine Learning on Massive High-Dimensional Vectors* (Université de Paris), *Security and Localization of IoT Devices* (Northeastern University), *High Speed Free Space*

Wireless Communications (Télécom ParisTech and KAUST), and *Arabic Dialect Identification* (Université de Montréal).

World-class resources. UM6P-CS taps state-of-the-art resources that enable innovative research and teaching. For instance, researchers have access to the powerful computing resources of the Benguerir Data Center (BDC), which has server rooms spanning 2,000 square meters, with a five-megawatt IT load. Besides, professors can use the services of the university's Digital Learning Lab to develop courses based on novel pedagogical approaches and share them with the community through the Digital Learning Platform (DLP). The BDC has been certified as a Tier III and Tier IV center by the Uptime Institute^k and the DLP, containing over 400 courses, was made accessible free-of-charge to thousands of students nationwide since the onset of the COVID crisis.^l

Turning ideas into reality. Students acquire interdisciplinary knowledge by developing and testing their ideas in real-scale experimental platforms such as living labs, experimental farms and mines, a green energy park, a smart building park, and a fab lab. The Technology Transfer Office supports university researchers in the commercialization of their inventions. Entrepreneurship is promoted within and outside the university with investment opportunities secured through venture capital firms and angel investors. 118 ideas have been supported so far through the pre-incubator programs Explorer and P-Curiosity Lab, in collaboration with MIT Sandbox. Ten projects are incubated by the U-Founders program, and 20 startups are supported by the Impulse accelerator program in collaboration with MassChallenge.

Innovation culture through location. UM6P-CS is located in the city of Ben Guerir, the first city in the African continent to participate in the Leadership in Energy and Environmental Design (LEED) Neighborhood Development certification process. This environment facilitates the develop-


^k <https://bit.ly/3qPU7Ht>

^l <https://bit.ly/37Ucm5F>

ment of technological innovations at the service of sustainable development. The university obtained a Silver level accreditation to the international standard Sustainability Tracking, Assessment & Rating System (STARS), awarded by the Association for the Advancement of Sustainability in Higher Education (AASHE).^m

Exchanging ideas. Prominent researchers from the Moroccan diaspora currently form the scientific advisory board at UM6P-CS. They are fully engaged in mentoring the next generation of researchers and supporting regional conferences. NETYSⁿ is one such conference that has allowed, since 2013, hundreds of students, who do not have the means to attend conferences overseas, publish their ideas, and get feedback from eminent researchers and scholars.

Conclusion

We presented four case studies of diverse universities across the Arab region, as well as some of the enabling factors that contributed to the creation of a thriving culture of innovation and discovery at the undergraduate level. We shared some better practice guidelines that helped stimulate undergraduate research with the aim of initiating further discussions about sparking innovation among undergraduate student populations elsewhere in the region, and worldwide. 

^m <https://bit.ly/3455WQe>
ⁿ netys.net

References

1. Anderson, R.D. Germany and the Humboldtian model. *European Universities from the Enlightenment to 1914*, Oxford University Press, 2004.
2. Baydoun, E. and Hillman, J. *Universities in Arab Countries: An Urgent Need For Change*. Springer International Publishing, 2018.
3. Islam, S.I. Arab women in science, technology, engineering and mathematics fields: The way forward. *World J. Education* 7, 6 (Nov. 29, 2017), p. 12; 10.5430/wje.v7n6p12.
4. Schimank, U. and Winnes, M. Beyond Humboldt? The relationship between teaching and research in European university systems. *Science and Public Policy* 27, 6 (2000), 397–408.
5. UNESCO Study Report On Financing Higher Education In Arab States, 2018; <https://en.unesco.org/sites/default/files/financing.pdf>

Slim Abdennadher, The German University in Cairo, Egypt.

Sherif G. Aly, The American University in Cairo, Egypt.

Joe Tekli, Lebanese American University, Lebanon.

Karima Echihabi, Mohammed VI Polytechnic University, Morocco.

© 2021 ACM 0001-0782/21/4

BY SEIF ELDAWLATLY, MOHAMED ABOUELHODA,
OMAR S. AL-KADI, TAKASHI GOJOBORI,
BORIS JANKOVIC, MOHAMAD KHALIL,
AHSAN H. KHANDOKER, AND AHMED MORSY

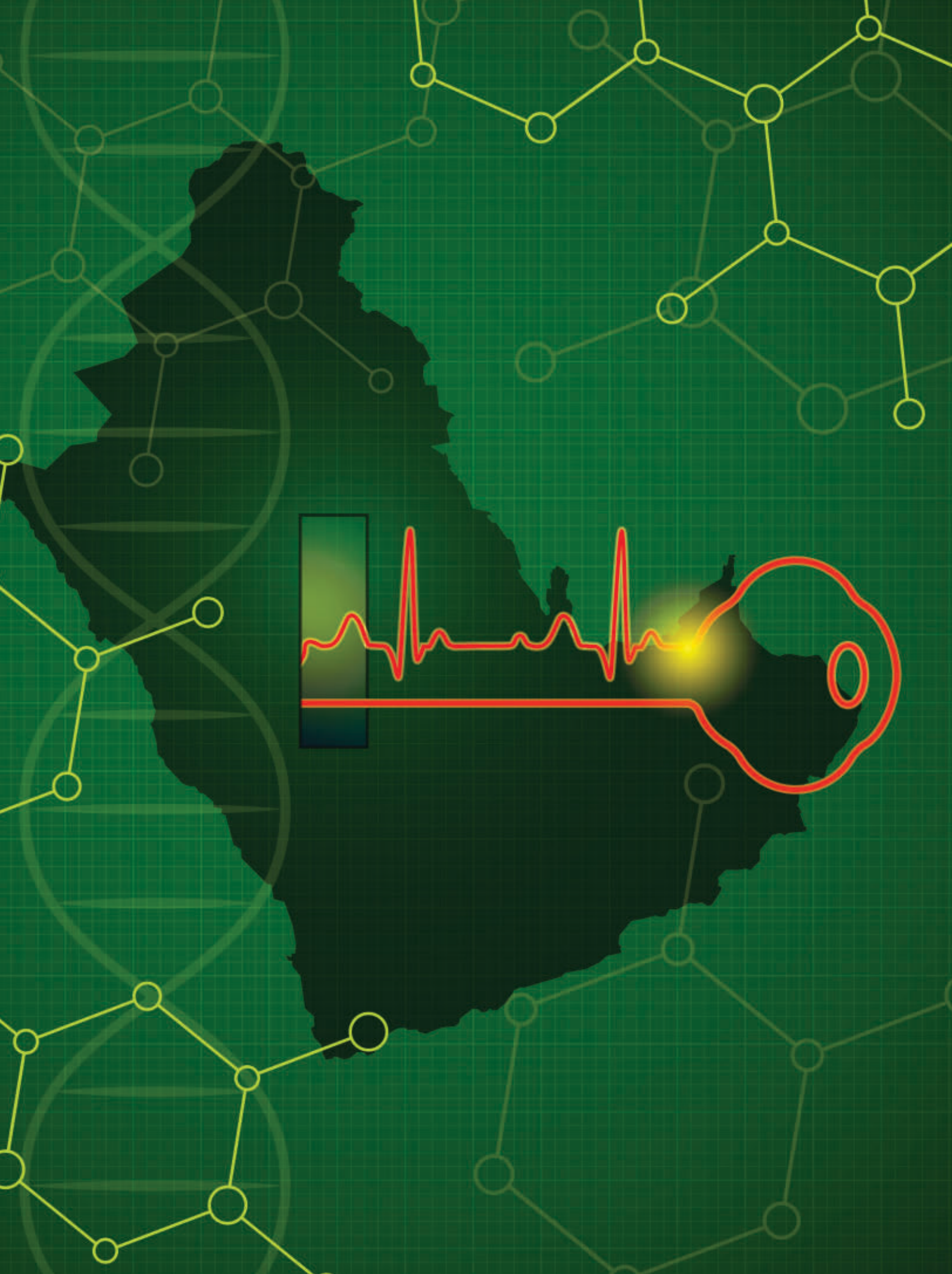
Biomedical Computing in the Arab World: Unlocking the Potential of a Growing Research Community

HEALTH CHALLENGES REPRESENT one of the long-standing issues in the Arab region that hinder its ability to develop. Prevalence of diseases such as cardiovascular diseases, liver cirrhosis and cancer among many others has contributed to the deteriorated health status across the region leading to lower life expectancy compared to other regions. For instance, the average life expectancy in the Arab world is approximately 70 years, which is at least 10 years lower than most high-income countries.² Among many directions of healthcare development across

the region, biomedical computing research represents one main arm of tackling health challenges. Advances in computational technologies have enabled the emergence of biomedical computing as one of the most influential research areas worldwide. In recent years, life sciences have witnessed an explosion in the volumes of biomedical data generated by high-throughput technologies and other sources. The enormity of volumes and interdependence in biomedical data pose great analytical challenges in the quest to infer deeply hidden knowledge buried under this complexity. As such, the biomedical computing research community in the Arab world has been actively contributing to the efforts that tackle long-standing biomedical challenges.

Research in biomedical computing in the region dates back to mid-1970s with the establishment of the Systems and Biomedical Engineering department at Cairo University in Egypt. Since then, the number of related programs has steadily increased and researchers from different disciplines have developed interest in biomedical computing applications. Despite the limited available resources, researchers from the Arab region have made over the years strong contributions to the rapid advances that occur in the field of biomedical computing. Recent successful efforts by researchers across the region have been evident in three broad areas of biomedical computing; namely, biomedical imaging, biomedical signal analysis and bioinformatics. These efforts have materialized in advancing a diverse spectrum of biomedical computing applications, as well as stimulating clear commercial interest.

This article sheds light on notable research efforts in the Arab world in each of the aforementioned areas of biomedical computing. It also demonstrates how this research



addresses healthcare issues in the region. A glimpse of the impact of research in this area in stimulating the budding culture of entrepreneurship and startup of new ventures across the region will be discussed. Finally, we introduce a vision for different avenues of development that could magnify the outcomes of the biomedical computing research community in the region.

Biomedical Imaging

Biomedical image analysis has always constituted a main line of biomedical computing research in the Arab world. Medical imaging modalities became widely available in the region in response to the strong prevalence of diseases that rely on imaging techniques for accurate diagnosis. This has made the case for biomedical image analysis research compelling, and many institutions in various countries in the region recognized this opportunity. Interest in biomedical image analysis research involving machine learning techniques has also become strong, fueling research on computer aided diagnosis and other predictive analytics techniques.

Early research in this area came from Cairo University (Cairo, Egypt) with the contributions led by Abou Bakr Youssef in tissue characterization using ultrasound images since the 1980s.¹⁵ Later on, a strong interest in magnetic resonance imaging (MRI) was developed, with an early focus on K-space and later on diffusion tensor imaging.¹¹ Ultrasound image analysis regained momentum with the work on elasticity and strain imaging. Other biomedical image analysis research that came out of Cairo University included localization of cardiac structures using MRI, identification of schizophrenic patients using functional-MRI,⁴ diagnosis of Alzheimer's disease using diffusion tensor images, skin lesion classification via optical images, designing insoles to address diabetic foot complications,¹² lung cancer diagnosis and prognosis evaluation, optical finger print recognition, and other image-based biometric identification techniques. Intelligent segmentation of cardiac structures

and evaluation of cardiac dynamics were also the research focus of two prominent research institutions in Egypt-Aswan Heart Center and Nile University.

Across the region, numerous research groups have published work in this area, utilizing various machine learning techniques. The work at the Computational Biomedical Imaging Lab at the University of Jordan (Amman, Jordan) is concerned with computer-aided diagnosis for improving tissue characterization and understanding of disease and tumor behavior.⁵ The lab also utilizes fractal analysis for classification of meningioma, the most common type of brain tumors, and segmentation of cardiac structures. These methods work either on a cellular level or tissue level.⁶ Research on segmentation of cardiac, breast, and brain images was also heavily published utilizing fuzzy logic methods (University of Mascara, Algeria), histogram-based techniques (United Arab Emirates University, Ajman, UAE), and various supervised learning techniques (King Abdelaziz University, Saudi Arabia). With the onset of the COVID-19 crisis, many researchers have also proposed methods for fast and accurate CT image segmentation, which is crucial to the diagnosis of COVID-19.²⁵ These research efforts, among many others, demonstrate the alignment of biomedical imaging research in the region with global healthcare issues.

Biomedical Signal Analysis

Biomedical signal analysis has risen as one of the key research areas, given the advances in the technology of recording different physiological signals from the human body. These signals can be utilized in diagnosing various diseases as well as modulating the function of different organs.

One example is the work of the Biomedical Signal Processing research group at Khalifa University (KU, Abu Dhabi, UAE) in the area of cardiovascular disease, which represents a leading cause of death in the region as well as worldwide. One line of work of the KU team is developing noninvasive fetal phono-

cardiogram as well as adult Electrocardiogram (ECG) signal processing techniques to prevent stillbirths and sudden cardiac deaths. The KU team successfully demonstrated the first proof of concept that a low-cost phonocardiogram sensor can detect fetal heart sounds and give a reliable estimation of the fetal heart rate and its variability, which were validated by simultaneously recorded fetal ECG signals.¹⁷ The work in this study resulted in a handheld fetal phonogram working prototype, which was validated in a number of healthy pregnancies. Additionally, the KU team has contributed to the worldwide research efforts to diagnose and predict cardiac arrhythmia complications. The KU team has developed a new device presenting a novel algorithm which was implemented in an application specific integrated circuit (ASIC) chip by using ECG signals to predict a heart attack long before its onset.⁹

Brain signal analysis represents one other notable research direction that is being pursued by multiple groups around the region. The LASTRE group at the Lebanese University (Tripoli, Lebanon) has been working on developing novel 'neuromarkers' to identify and characterize networks associated with cognitive deficits in patients, particularly at early stage. It is recognized that neurological pathologies, such as Alzheimer's disease (AD), are caused by alterations in these brain networks. In this context, the LASTRE group investigated dynamic topological changes of AD networks in terms of brain network segregation and integration.¹⁴ In their analysis, functional brain networks were reconstructed from brain Electroencephalography (EEG) activity in different frequency bands. The achieved results revealed that networks in AD patients are characterized by less integration and higher segregation compared to networks reconstructed from healthy subjects. This could complement current AD diagnostic metrics, especially at early stages of the disease.

In addition, multiple research groups in the region have contributed to the development of a plethora of brain signal analysis techniques


for a variety of applications. One study from the Biomedical Engineering group at the American University of Sharjah (AUS, Sharjah, UAE) has proposed a technique to assess the mental capacity to preserve attention for long durations.⁸ Their technique was able to monitor changes in the communication patterns among different brain regions with reduced attention. Another research direction that is being pursued in this area is driving brain activity through artificial stimulation, which has shown its merits in compensating for different types of disability. One example is the work of the Biomedical and Neuro Engineering Laboratory (BNEL) at Ain Shams University (ASU, Cairo, Egypt) to develop visual prostheses for restoring vision to the blind. They have proposed a novel Kalman filter-based technique to automatically tune electrical stimulation delivered through visual prostheses devices which has been shown to evoke neuronal activity similar to natural responses.¹³ Taken together, biomedical signal analysis research in the region has resulted in influential and diverse contributions that aim at resolving multiple technical challenges in the field and at addressing several population health issues.

Bioinformatics


The Arab world has a vast and versatile environment inhabited by large populations having interesting structure and dynamics. This has led to multiple bioinformatics research efforts that employ high-performance computational methods to tackle hereditary diseases prevalent in the region.

There have been multiple efforts around the region to develop national genome programs. One prominent example is the Saudi Human Genome Program, which has the capacity of 100,000 samples. The project focuses on unraveling the mutations responsible for inherited disorders in the Saudi/Arab population. The project, which was started in 2014, is recognized as one of the top 10 human genome projects worldwide and is conducted by the King Abdulaziz City for

Science and Technology (KACST, Riyadh, Saudi Arabia).²⁴ The findings of the project have paved the way for the implementation of precision medicine in Saudi Arabia. The project could identify many new mutations related to disease. It has also reclassified, as benign, many variants, previously thought to be responsible for disease because of limited data.³ King Faisal Specialist Hospital and Research Center (KFSHRC, Riyadh, Saudi Arabia) has launched the Clinical Genomics transformation project, which leverages genomic testing for clinical practice. It has already contributed to the successful diagnosis and treatment of thousands of patients with inherited disorders as well as cancer.²⁰ The Qatar genome project, which is one of the early projects in the region, was launched to study and isolate genetic characteristics of the local population. The first phase was completed successfully, with the analysis of 20,000 genomes.²³ The current phase of the project targets 10% of the population, and subsequent phases have an ambitious goal of completing the analysis of 350,000 genomes. The Emirati project has also completed the characterization of 1,000 individual genomes,⁷ with aspirations to eventually cover the entire population of the country. Very recently, Egypt also announced the starting of the Egyptian Genome Program targeting 100,000 samples, in addition to a number of studies to analyze cohorts of patients. Although of limited scale, these studies shed light on the genetic bases of certain diseases in the Egyptian population.²² North African Arab countries Tunisia, Morocco, Sudan, and Egypt have also joined the African genome project H3Africa, which is an Africa-wide initiative aiming at studying genetic variations in the continent. Genetic studies for population characterization and disease studies have also been conducted by researchers from Lebanon, Tunisia, Kuwait, and Bahrain.¹ Researchers from Qatar Computing Research Institute (QCRI) have demonstrated the use of machine learning tech-



Biomedical signal analysis has risen as one of the key research areas, given the advances in the technology of recording different physiological signals from the human body.



There have been multiple efforts around the region to develop national genome programs.

niques in characterizing molecular interactions of cancer subtypes.¹⁹ In addition to medical research, there was also a number of genome projects of regional interest that were successfully completed. These included the Date Palm genome projects in Saudi Arabia and Qatar, and the Egyptian Buffalo genome projects. These projects serve as a foundation for improving the traits of these plants and animals to optimize food production.

Recognizing the importance of this research area, the Computational Bioscience Research Center (CBRC) was established at the inception of King Abdullah University of Science and Technology (KAUST, Thuwal, Saudi Arabia) in 2009. The interdisciplinary nature of research requires designs of integrated computational and experimental methods and tools for use in life sciences and biotechnology development. The computational and experimental facilities at KAUST enables CBRC to achieve such a synergistic approach. In collaboration with local health care providers, CBRC research is particularly focused on population and comparative genetics and genome-wide association studies. One research thrust at CBRC is in the field of metagenomics, which is useful in many applications, such as soil management, bioprospecting for industrially relevant compounds, detection of novel genes and proteins, medical diagnostics and epidemiology, oil industry and many others. In particular, CBRC utilizes metagenomics for bio-prospecting of cellulase for exploring novel enzymatic genes in the Red Sea, including those for bio-fuel development.¹⁰ CBRC also conducts research in structural biology; proteins structure analysis, prediction, and engineering; and cellular signaling. A related work aims for the development of computational tools and resources for designing efficient microbial cell factories.²¹ These are microorganisms whose metabolic processes are altered, for example through gene editing methods, in such a way that they increase production of chemical compounds that may be of industrial or pharma-

ceutical interest. These examples show the contributions of bioinformatics researchers in the region to enhance the quality of life of millions of people around the region.

Entrepreneurial Activities

Built upon the success demonstrated in different biomedical computing tracks, the Arab region has witnessed in the past few years a strong momentum for entrepreneurial activities in many sectors. While biomedical computing research in the region did not lean toward the translational side that transforms basic research outcomes to products and tools, we have witnessed a few cases of patents resulting from research and fueling the birth or the growth of startups. For example, the work of the Biomedical Signal Processing research group at Khalifa University raised a significant commercial interest that resulted in a UAE-based startup company licensed to commercialize a phonogram technology for home monitoring of fetal well-being.¹⁶ Another example is in the field of smart radiology and ultrasound imaging analysis which materialized into an Egypt-based startup.¹⁸ Moreover, several graduates of biomedical engineering programs, medical schools, and graduates of other life science related programs have started up their own health-tech ventures. Some examples of these cases cater to radiology computer-aided diagnosis (Intexil, <https://www.intixel.com/about-us/>), smart radiology workflow (Dileny Tech, <http://www.dilenytech.com/>), radiology reporting (Rology, <https://rology.health/>), cardiac events monitoring (BioBusiness, <http://www.biobusiness-eg.com/>), remote patient monitoring (Pulse, <http://pulse-eg.com/>), diabetic eye care (<http://www.almouneer.com/>), bioinformatics (Proteinea, <https://proteinea.com/>), and early diagnosis of cancer (Prognica, <http://www.prognica.com/>). The emergence of these entrepreneurial activities is driven by both increased interest from investors and expanded awareness of researchers and graduates of relevant programs.

Regional Conferences

There are three main regular biomedical engineering conferences in the Arab world. The first to get started was the Cairo International Biomedical Engineering Conference (CIBEC), which is held in Cairo in December every other year since 2002. The conference with the highest outreach is the Middle East Conference on Biomedical Engineering (MECBME), as it rotates among different countries in the region, including UAE, Jordan, and Tunisia. The International Conference on Advances in Biomedical Engineering (ICAMBE) is held regularly in Lebanon, and typically enjoys a high level of attendance from many Franco-Arab researchers living abroad, in addition to researchers from around the region. An important observation regarding these three conferences is the high level of coordination among their organizers in order to avoid time overlap and to increase geographic coverage. These conferences have succeeded over the years in bringing together biomedical computing researchers from across the region.

Conclusion and Future Perspectives

The biomedical computing research community in the Arab world has demonstrated valuable success over the years in multiple facets of research. In this article, we focused on three biomedical computing tracks: biomedical imaging, biomedical signal analysis and bioinformatics; outlining examples of outstanding research being carried out in the region in each track. In addition, we discussed the success of the biomedical computing community in the Arab World in establishing entrepreneurial activities. Yet, multiple aspects of development still need to be pursued. What the region lacks fundamentally is initiatives to build and publish datasets based on populations from countries in the region itself. In addition, cross-country collaboration is scarce (see for example Abdelhedi et al.¹), although if activated it can have a positive impact on both the quality and quantity of biomedical computing research out-

put. The regional conferences outlined in this article could represent one valuable opportunity to initiate the pursued cross-country collaborations. Another potential aspect of development is establishing animal research labs, which can provide a critically needed dimension of biomedical computing research. Few labs have started to adopt this approach including, for example, the Biomedical and Neuro Engineering Laboratory at Ain Shams University and the Biomimetics Engineering Laboratory at the American University of Beirut (Beirut, Lebanon). However, providing the financial and logistical resources needed for the wide-scale establishment of animal research labs is still lacking. Despite all challenges facing researchers, the promising success of biomedical computing researchers within the Arab region as demonstrated in this article indicates that unlocking their full potential could significantly enhance the health and wellness of millions of people across the region as well as worldwide. 

References

1. Abdelhedi, R. et al. Characterization of drug-metabolizing enzymes CYP2C9, CYP2C19 polymorphisms in Tunisian, Kuwaiti and Bahraini populations. *Journal of Genetics* 94 (2015), 765–770.
2. Abdelmoneim, A.O. and Alharahsheh, S.T. Family home caregivers for old persons in the Arab region: perceived challenges and policy implications. *Open Journal of Social Sciences* 4 (2016), 151–164.
3. Abouelhoda, M. et al. Revisiting the morbid genome of Mendelian disorders. *Genome Biology* 17 (2016), 235.
4. Algunaïd, R.R. et al. Schizophrenic patient identification using graph-theoretic features of resting-state fMRI data. *Biomedical Signal Processing and Control* 43 (2018), 289–299.
5. Al-Kadi, O.S. A multiresolution clinical decision support system based on fractal model design for classification of histological brain tumours. *Computerized Medical Imaging and Graphics* 41 (2015), 67–79.
6. Al-Kadi, O.S. Spatio-temporal segmentation in 3D echocardiographic sequences using fractional Brownian motion. *IEEE Transactions on Biomedical Engineering* 67 (2020), 2286–2296.
7. AlSafar, H.S. et al. Introducing the first whole genomes of nationals from the United Arab Emirates. *Scientific Reports* 9 (2019), 1–15.
8. Al-Shargie, F.M. et al. EEG-based semantic vigilance level classification using directed connectivity patterns and graph theory analysis. *IEEE Access* 8 (2020), 115941–115956.
9. Bayasi, N. et al. Low-power ECG-based processor for predicting ventricular arrhythmia. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* 24 (2015), 1962–1974.
10. Behzad, H. et al. Metagenomic studies of the Red Sea. *Gene* 576 (2016), 717–723.
11. Gabr, R.E. et al. Deconvolution-interpolation gridding (DING): Accurate reconstruction for arbitrary k-space trajectories. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 56 (2006), 1182–1191.
12. Ibrahim, M. et al. A pilot study to assess the effectiveness of orthotic insoles on the reduction of plantar soft tissue strain. *Clinical Biomechanics* 28 (2013), 68–72.
13. Jawwad, A. et al. Modulating lateral geniculate nucleus neuronal firing for visual prostheses: A Kalman filter-based strategy. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25 (2017), 1917–1927.
14. Kabbara, A. et al. Reduced integration and improved segregation of functional brain networks in Alzheimer's disease. *Journal of Neural Engineering* 15 (2018), 026023.
15. Kadah, Y.M. et al. Classification algorithms for quantitative tissue characterization of diffuse liver disease from ultrasound images. *IEEE Transactions on Medical Imaging* 15 (1996), 466–478.
16. Khandoker, A. Low cost fetal phonocardiogram. Google Patents, 2018.
17. Khandoker, A. et al. Validation of beat by beat fetal heart signals acquired from four-channel fetal phonocardiogram with fetal electrocardiogram in healthy late pregnancy. *Scientific Reports* 8 (2018), 1–11.
18. Mahmoud, A.M.E. and Ali, M.T.M. Method and apparatus to measure tissue displacement and strain. Google Patents, 2020.
19. Mall, R. et al. Detection of statistically significant network changes in complex biological networks. *BMC Systems Biology* 11 (2017), 32.
20. Monies, D. et al. Lessons learned from large-scale, first-tier clinical exome sequencing in a highly consanguineous population. *The American Journal of Human Genetics* 104 (2019), 1182–1201.
21. Motwalli, O. et al. In silico screening for candidate chassis strains of free fatty acid-producing cyanobacteria. *BMC Genomics* 18 (2017), 33.
22. Nassar, A. et al. Targeted next generation sequencing identifies somatic mutations in a cohort of Egyptian breast cancer patients. *Journal of Advanced Research* 24 (2020), 149–157.
23. Qoronfleh, M.W. et al. The future of medicine, healthcare innovation through precision medicine: Policy case study of Qatar. *Life Sciences, Society and Policy* 16 (2020), 1–20.
24. S.G.P. Team. The Saudi Human Genome Program: An oasis in the desert of Arab medicine is providing clues to genetic disease. *IEEE Pulse* 6 (2015), 22–26.
25. Zhou, L. et al. A rapid, accurate and machine-agnostic segmentation and quantification method for CT-based COVID-19 diagnosis. *IEEE Transactions on Medical Imaging* 39 (2020), 2638–2652.

Seif Eldawlatly, Computer and Systems Engineering Department, Faculty of Engineering, Ain Shams University, Cairo, Egypt, and Faculty of Media Engineering and Technology, German University in Cairo, Cairo, Egypt.

Mohamed Abouelhoda, Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt, and, King Faisal Specialist Hospital and Research Center, and Saudi Human Genome Program, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia.

Omar S. Al-Kadi, King Abdullah II School for Information Technology, University of Jordan, Amman, Jordan.

Takashi Gojobori, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Boris Jankovic, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Mohamad Khalil, Faculty of Engineering, Lebanese University, Tripoli, Lebanon.

Ahsan H. Khandoker, Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates.

Ahmed Morsy, Systems and Biomedical Engineering Department, Faculty of Engineering, Cairo University, Giza, Egypt.

Copyright held by authors/owners.

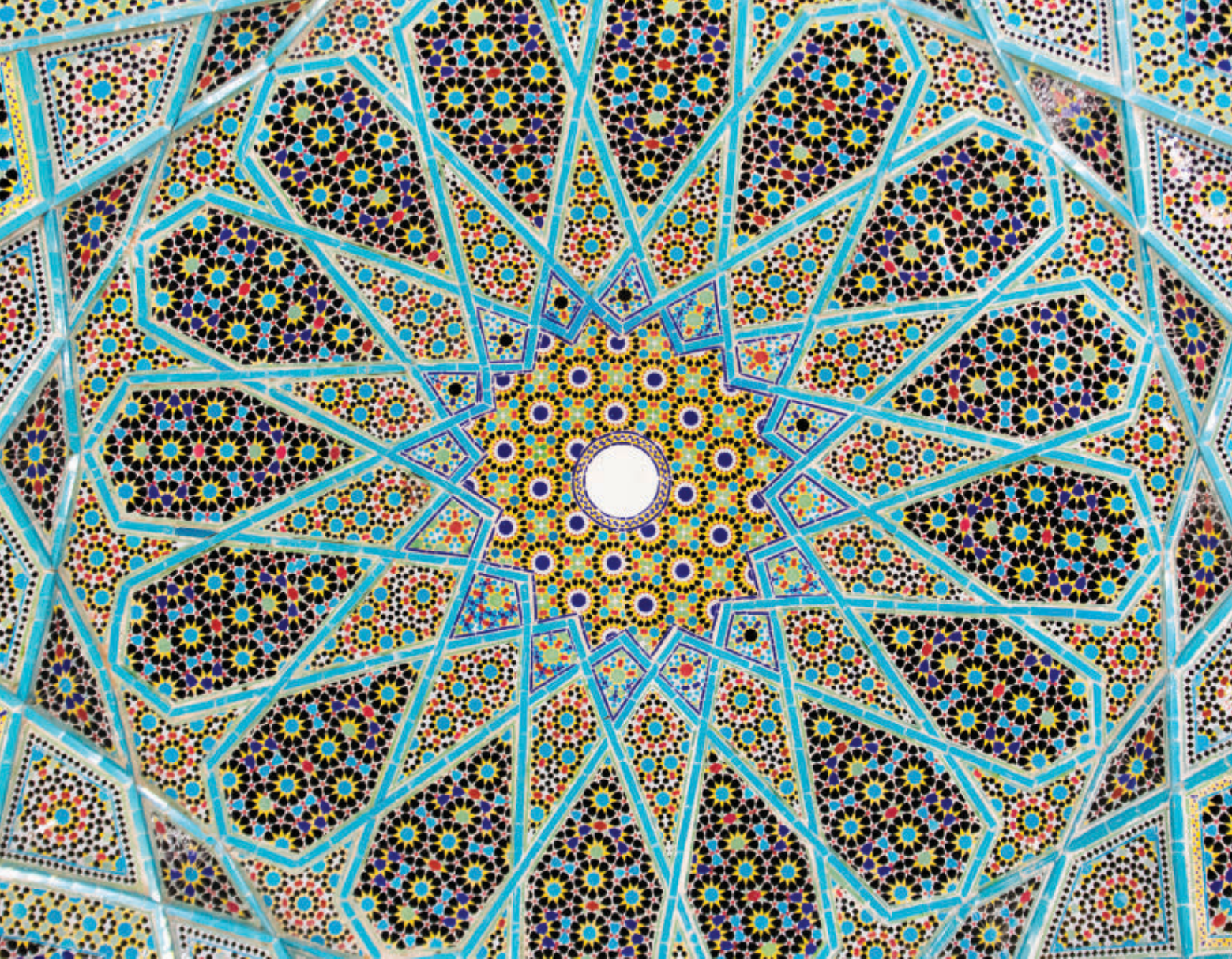
BY BASEM SHIHADA, TAMER ELBATT, AHMED ELTAWIL,
MOHAMMAD MANSOUR, ESSAID SABIR, SLIM REKHIS,
AND SANAA SHARAFEDDINE

Networking Research for the Arab World: From Regional Initiatives to Potential Global Impact

THE ARAB REGION, composed of 22 countries spanning Asia and Africa, opens ample room for communications and networking innovations and services and contributes to the critical mass of the global networking innovation. While the Arab world is considered an emerging market for communications and networking services, the rate of adoption is outpacing the global average. In fact, as of 2019, the mobile Internet penetration stands at 67.2% in the Arab world, as opposed to a global average of 56.5%.¹² Furthermore, multiple countries in the region are either building

new infrastructure or developing existing infrastructure at an unprecedented pace. Examples include, Neom city in Saudi Arabia, the new administrative capital in Egypt, as well as the Smart Dubai 2021 project in the United Arab Emirates (UAE), among others. This provides a unique opportunity to fuse multiple advanced networking technologies as an integral part of the infrastructure design phase and not just as an afterthought.

Among a number of emerging communications and networking technologies, wireless and mobile technologies are of paramount importance and have become a key enabler for a multitude of services in our daily lives. Nowadays, mobile broadband technologies offer ubiquitous access to novel services and Internet access to billions of people around the world. The Arab world is no exception, considering the research work done in the region for advancing wireless systems and deployments. Over the course of each decade starting 1980, the world has witnessed four generations of cellular networks, each introducing a specific set of technologies and a portfolio of supported new use cases. As a 4G standard, Long Term Evolution (LTE) has become the most successful mobile wireless broadband technology, serving billions of users while handling a wide range of applications. Despite some disparity in regional availability of LTE/4G, the telecommunications industry has already started rolling out the fifth generation (5G) of mobile communications since late 2018, bringing a significantly wider range of new use cases and unprecedented capabilities. Example 5G technologies that enable this vision include, but are not limited to, pervasive artificial intelligence, beyond THz-band communications, intelligent reconfigurable surfaces, and Internet of Space Things (IoST), among others. The latter is instrumental in extending seamless connectivity and mobile services to rural, subma-



rine, and hard-to-reach areas, through an integrated system of CubeSats, tethered balloons, and unmanned aerial vehicles (UAVs). To match those needs, networks have evolved in numerous ways, including the growing role of artificial intelligence and machine learning (AI/ML) in designing and optimizing networks and extensions to airborne and underwater networking infrastructure, as will be presented in this article. In addition to technology breakthroughs, novel rollout roadmap, market readiness, education, and research opportunities also emerge all over the globe.¹

There is a tremendous amount of effort going into networking research at Arab academic and research institutions. Those efforts are not only motivated by the unique needs of the Arab world but also have a potential for global impact due to the similarity of networking research challenges in

different parts of the world. In light of this promising trend and technological progress of networking in the Arab world, we unveil in this article a sample of relevant research activities. We have organized the article into five themes that symbolize collaborative efforts across the Arab world for solving challenging and intriguing networking research problems that are key to the region's sustainable economic development and prosperity.

Machine Learning and Edge Computing in the Newly Developed Arab Smart Cities

In today's complex communication systems, not only is the parameter space large, but it is also time varying, and more often than not, involves objective functions that are highly non-linear and/or of polynomial/exponential complexity. ML offers a solution for setting and continuously updating

key network parameters based on actual usage patterns. ML models that can learn from prior experience/data using automatic feature extraction to alleviate the need for tedious manual extraction.

In a project jointly initiated by King Abdullah University of Science and Technology (KAUST) and the American University in Beirut (AUB), the researchers plan to review how AI/ML can be used for a variety of wireless design problems, including channel modeling and estimation, channel encoding and detection, power control and beamforming, and network optimization and resource allocation.⁹ In addition, opportunities and challenges are explored when integrating AI/ML approaches with new infrastructure deployment which lacks historical usage patterns and, hence, implies alternative approaches such as transfer learning and online training

and inference. The systems mentioned in this article have a great potential for planning reliable wireless systems for future Arab smart cities. These smart cities are special of its kind as they have the advantages of being planned and constructed on a new landscape, integrate the latest-edge technologies, consider the environmental effects, and operate on clean energy.

On a related front in Egypt, an ongoing research project funded by the Information Technology Industry Development Agency (ITIDA) and led by ElBatt at the American University in Cairo (AUC), in collaboration with the industrial partner IoTBlue,^a focuses on leveraging edge computing, jointly with cloud computing, towards distributed machine learning and scalable Internet of Things (IoT). The project involves the research, design and development of a prototype for a multi-tier computing system, spanning the edge, gateway and cloud tiers, with small-scale, short-term learning carried out at the two edge tiers while large-scale, long-term analytics at the cloud. The major contribution of this work is twofold: a novel machine learning model; distributed across devices in a single tier and across multiple tiers and, achieving minimal communications overhead while attaining similar or close prediction accuracy to the centralized ML baseline. It gives rise to a multitude of interesting research problems ranging from architecting the multi-tier computing system, distributed machine learning, intra-edge tier and across tiers, hosting heterogeneous sensors and IoT platforms, wireless networking and edge computing. The proposed system finds wide application in diverse IoT verticals like health, intelligent transportation and smart cities, being highly relevant to the new administrative capital, currently under development, as well as Cairo which is considered the most populous city in the Arab world and Africa and the sixth populous city in the world.

Finally, a relevant research thrust at Carnegie Mellon University in Qatar

lead by Harras targets intelligent resource allocation for the mobile cloud, in particular FemtoCloud and edge computing. FemtoCloud has a vital role in facilitating ultra-reliable and low-latency traffic and decision making for sensitive smart-city applications. For instance, RAMOS⁵ is a resource-aware multi-objective system for edge computing. The system prototype achieves up to 40% completion time improvement under latency minimization mode and up to 30% more energy-efficiency under the energy-efficient mode. The team also developed computational offloading schemes that maximize the lifetime of the ensemble of mobile devices where they consider a mobile network while no device has depleted its battery. They also demonstrated the effectiveness of the computation offloading system that contributed to extending the lifetime of a mobile device cloud.

Flying UAVs for Smart Management in the Arab Cities

Flying networks are envisioned to play a vital role in city and disaster management through disaster prediction, response, recovery, as well as mitigation. They are also considered as one of the most promising technologies for disaster management according to recent Red Cross and UN reports.¹⁴ They present themselves as aerial base stations equipped with intrinsic communications, computing, and sensing capabilities. In particular, unmanned aerial vehicles (UAVs) have attracted strong interest from, both, research and industry communities owing to their agility, flexibility, and low cost. In terms of communications and connectivity, UAVs are deployed to provide immediate network infrastructure recovery for potential survivors and first responders, yielding timely emergency response.

A research project carried out by the Lebanese American University (LAU) deploys a swarm of UAVs that are autonomously distributed in 3D space to provide the needed network coverage over a given disaster scene and adapts its location dynamically to accommodate user mobility using low-complexity, real-time algorithms. This research is extended to develop a novel approach that allows rescue

teams locate victims based on the received signal strength indicator from their mobile devices. The proposed localization approach caters to disaster scenes with heterogeneous structure, allowing areas with a higher degree of damage or population density to be better served with a finer level of accuracy. To do so, UAVs sniff wireless signals from the victims' mobile devices to accurately determine their locations using optimized trajectories while accounting for channel variability. Low complexity algorithms have been developed and validated using an experimental testbed under realistic conditions to demonstrate the feasibility and effectiveness of UAV-based flying networks during emergency relief and rescue operations. Moreover, the UAV can be augmented with a reinforcement learning agent to search for victims and continuously improve the localization precision as long as the UAV's energy is not drained.

On another front in the UAE, the center for autonomous and robotic systems at Khalifa University, along with the internationally recognized MBZIRC robotic challenge partners¹³ are focusing on next generation intelligent robotics. This includes low altitude visual tracking UAVs that utilize deep learning. They run several projects for smart city management using UAV autonomous swarms and environmental imagery.¹⁰ As Dubai announced the Smart Dubai initiative, local research institutions devoted much efforts toward utilizing UAVs to cater to this ambitious initiative. Dubai have deployed UAVs for Geospatial and surveying activities. They also deployed it for civil security control, traffic management, agriculture projects and environmental management.

Underwater Internet Using Wireless Optical Networks for Red Sea Coral Reef Monitoring

Underwater wireless communications play an important role in environmental monitoring, underwater exploration, and scientific data collection. In this arena, KAUST is leading multiple research projects relevant to coral reef monitoring in the Red Sea. This environmental and marine life regional

^a IoTBlue (<http://www.iotblue.net>) is actively participating in a number of projects related to automation and smart cities in Egypt and the region.

need is motivated by evidence that coral reefs are facing growing challenges which call for extensive multi-disciplinary research to better understand the phenomena and address it. State-of-the-art submarine exploration missions typically demand efficient, flexible, and high data rate communications to access biological data collected by divers and unmanned vehicles.

Conventional radio frequency (RF) and acoustic communications critically suffer from severe attenuation of RF signals in water and low data rates (100bps to 100kbps), for acoustic waves, due to the low propagation speed (1500m/s) and large latency. On the other hand, underwater optical communications, based on diffuse (LED) and collimated (Laser) light sources, has been explored as a promising solution to support much higher data rates over ranges up to several tens of meters, thanks to their cost-effectiveness and low power consumption. In light of this, the research project of Underwater Internet Using Wireless Optical Networks led by Alouini, Ooi, and Shihada provides high data rate, low-power underwater communications that can overcome the aforementioned channel impairments and has practical relevance.

This project successfully characterized the statistics of fading for underwater optical wireless communication (UOWC) channels and analyzed the system performance. Furthermore, it developed a novel high-speed gallium nitride (GaN)-based laser diode that has a set of unique traits tailored for the Red Sea. Also, the project developed a low-cost, energy-efficient transceiver for UOWC systems supporting high data rates (up to 1Gbps over ranges up to 120 meters). Finally, the team built a prototype and performed system testing and debugging to ensure smooth, reliable live video streaming using underwater laser diodes. The research team has further extended the system to the experimental stage with a breakthrough demo, for the first time, of a bi-directional UOWC system capable of transmitting an ultra-high-definition (UHD) quality, real-time video over 4.5m. Following this break-

through, a low-power and compact UOWC system, coined Aqua-Fi,¹¹ has been proposed and deployed by the same research team. Aqua-Fi is an underwater wireless Internet solution hosting an Internet bridge, a transmission modulator onto an optical carrier, a transmitter equipped with projection optics and beam steering elements, a detector, and a signal processing unit. Aqua-Fi is a cost-effective solution using low power, off-the-shelf components (for example, LEDs and lasers) to achieve at least 1Mbps underwater data rates.

The successful implementation of Aqua-Fi could significantly advance work in coral reef ecosystems, in general, by facilitating the real-time transfer of continuous data streams. Numerous applications, such as live feeds from monitoring equipment and other sensors, could provide management entities with dynamic information critical to their decisions. Also, Aqua-Fi facilitates communications and data sharing among divers, bringing marine science into the 21st century. Aqua-Fi has brought the attention of both the research and the industrial communities. For example, as of July 19, 2020, Yahoo (71.4 million read), TechCrunch (12.4 million), and Hindu (13.5 million). There are also several published video interviews, podcasts, and many others.^b

5G Research and Deployment Initiatives in North Africa


Over the past three decades, wireless technologies have experienced exponential growth, fueled by breakthroughs in semiconductors, thanks to Moore's Law, and information and communications theory. Apart from the currently launched 5G communication systems, beyond 5G (B5G) systems promise to reinvent entire industry sectors, by creating new use cases and models such as massive IoT connectivity, augmented reality, virtual reality (AR/VR), Vehicular-to-Anything (V2X), among others. These new use cases demand network support with large degrees of freedom. Furthermore, the vision for the sixth

^b Aqua-Fi Project <https://www.shihada.com/node/141>, and <https://github.com/CBinda-house/AquaFi>




There is a tremendous amount of effort going into networking research at Arab academic and research institutions.





In today's complex communication systems, not only is the parameter space large, but it is also time varying.



generation (6G) wireless networks is based on various breakthrough technologies expected to realize global connectivity across vertical industries with a plethora of innovative applications and services.⁴

Multiple research groups in Egypt are actively working on various aspects of 5G communications and networking research ranging from the network infrastructure to new application scenarios and emerging services. Those efforts include novel wireless technologies, mobile and edge computing and wireless test-beds for education and research. The importance of wireless and mobile networking technologies, in general, and 5G technologies to Egypt and the Arab world stems from: the ease of deployment, maintenance and upgrade in dense population cities, like Cairo and providing reliable Internet connectivity to remote, underserved areas, especially for remote health monitoring, learning and working, for example, at the time of COVID-19 lockdown. Among the prominent networking research activities in Egypt are, most notably, the ongoing work and collaboration of research groups at Cairo University, Alexandria University and Egypt Japan University of Science and Technology (E-JUST), The American University in Cairo (AUC) and the Wireless Intelligent Networks Center (WINC), Nile University. A significant body of research has been developed over the past few years on 5G networks, with publications in high-impact journals,^{2,4,6} through collaborative projects among researchers at the Cairo University, AUC and Nile University, in addition to Qatar University and Sabanci University, Turkey. First, the design and optimization of energy-efficient wireless networks and energy harvesting 5G networks toward reducing the carbon footprint in dense Arab cities, like Cairo. This among other networking research efforts hinge on the abundance of renewable energy sources in most Arab countries, for example, solar, wind, and RF interference.² Second, optimization of cognitive radio networks toward efficient spectrum utilization, which is a global research challenge with an

impact that goes far beyond the Arab world.

Caching at the wireless network edge,⁶ with potential applications in wide-scale content distribution to reduce the volume of download traffic at rush hours, for example, remote learning settings more common nowadays. Another research thrust at Cairo University has focused on 4G/5G interference management catering to dense population cities, for example, metropolitan areas.⁷

Finally, heading west in North Africa, namely to the “Maghreb” region (the western end of the Arab world), researchers with Hassan II University of Casablanca in Morocco and University of Carthage in Tunisia propose to introduce a set of best practices of 5G networks, in terms of deployment and spectrum management. Among multiple 5G configurations, two deployment models have been standardized to match market requirements and need to be expatiated on. The first deployment approach is Non-Standalone (NSA) where the 5G New Radio (NR) is connected to, and controlled by, the existing 4G core network and only supports Enhanced Mobile Broadband services. On the other hand, the Standalone (SA) approach fully exploits the capabilities of the 5G NR and the 5G core and allows to support Massive Machine Type Communications and ultra-reliable and low latency communication (URLLC) services.

Morocco was expecting to launch its first 5G network in late 2020. However, due to the COVID-19 outbreak, the 5G rollout has been postponed to 2021 or early 2022. In early 2019, Maroc Telecom (IAM) had teamed up with Ericsson to showcase a live 5G demonstration at IAM's headquarters in Rabat. This pilot staged several 5G use-cases demonstrating very high speeds reaching up to 25.8 Gbps. Indeed, Orange Morocco and Huawei have performed several 5G trials in March 2019. INWI is the first Moroccan operator to have signed an agreement with Huawei. Since mid-2019, they succeeded in deploying many pre-commercial 5G pilot projects supporting numerous 5G use-cases. For a smooth rollout, policymakers should reduce the

regulatory costs and fees, and provide affordable 5G handsets.

Based on the collaborative research effort across Morocco and Tunisia, it is observed that the Maghreb region has a strongly emerging 5G community aiming to facilitate the worldwide harmonization of research and best practices for the deployment of viable user scenarios within the global 5G ecosystem, built-in security and privacy by design in 5G, and explore the different approaches to reach efficient IP convergence.

Limited Resources and Remote Learning Support During COVID-19

Traditionally, networking protocols are validated using simulations, emulation, or testbeds. Simulations can provide large-scale evaluation but may not reflect reality. Emulation and implementation of real testbeds may be limited to the available resources. An interesting project at the Wireless Research and the Smart Critical Infrastructure centers at Alexandria University, Egypt is the WiPi Project.³ WiPi aims at providing a low-cost remotely accessible wireless networking testbed, a requirement for realistically validating networking protocols. It allows researchers and practitioners to run large-scale wireless experiments remotely. The testbed is realized using low-cost Raspberry Pi nodes as the computing resources and employs concepts such as resource pooling, node virtualization, and federation to maximize resource utilization and reduce the overall testbed cost. This not only allows for larger testbeds but also helps reach a larger pool of users, especially in these days of COVID-19 where remotely accessing education and research facilities is a priority. The testbed started as a single testbed in Alexandria University and is now federated to include Cairo University, AUC, E-JUST, and Assiut University. Users may run their experiments across nodes in the different sites concurrently and the testbed smoothly handles distributing the experiment on the testbed nodes to achieve a variety of goals.

Several research efforts highlighted the increasing evidence of viral spread due to aerosol transmission, where small particles can be trans-

ferred through the aerosol channel for long-distance causing infection. Recently, a research initiative from KAUST led by Alouini and Shihada have proposed a novel research direction called “Communication via Breath,” which visualizes the viral spread mechanism as a piece of transmitted information through a molecular communication channel.⁸ The proposed “Communication via Breath” concept utilizes the breath as a source of messages, where several bio-information can be transmitted through molecular aerosol channel. The information is carried by several biological entities such as pathogens and VOCs, which also act as health biomarkers. Bringing such a biological problem under the molecular communication umbrella enables researchers to apply mathematical methodologies, approaches, and analysis tools from information and communication technology research to model, analyze, study and deal with the problem of viral aerosol spread. Interestingly, the new visionary research attracted molecular communication researchers to conduct research that can support our concept. Notably, some algorithms are developed to estimate the transmission distance, and some model is proposed based on fluid dynamics. Indeed, the novel “Communication via Breath” research is related to the progress in three areas, information and communication technology, fluid dynamics, and molecular biology.⁸

Conclusion

In this article, we presented prominent networking research activities under five major research themes highly relevant to the Arab world. The prime objective of this article is to draw the attention of the international networking community and raise awareness about the imperative contributions of world-class research coming out of this part of the world. In addition, we highlight the key role networking is playing as part of the Digital Transformation initiatives taking place around the Arab world. Based on the outlined research efforts with regional focus yet hosting challenges common to other parts of the world, we envision ample room for a

potential global impact via leveraging key insights, enabling technologies, lessons learned and major findings of this research in other parts of the world. This article should also open venues for intra- and inter-region collaboration opportunities in the vibrant area of networking and related verticals. 

References

1. Arroub, A. et al. A literature review on smart cities: Paradigms opportunities and open problems. In *Proceedings of the 2016 Intern. Conf. on Wireless Networks and Mobile Communications* (Oct. 2016), 26–29.
2. Ashour, M. et al. Energy-aware cooperative wireless networks with multiple cognitive users. *IEEE Transactions on Communications* 64, 8 (Aug. 2016), 3233–3245.
3. Attaby, A. et al. Wipi: A low-cost large-scale remotely accessible network testbed. *IEEE Access* 7 (2019), 167795–167814.
4. Dang, S. et al. What should 6G be? *Nature Electronics* 3, 1 (2020), 20–29.
5. Gedawy, H.K. et al. RAMOS: A resource-aware multi-objective system for edge computing. *IEEE Transactions on Mobile Computing*, 2020.
6. Girgis, A. Fundamental limits of memory-latency trade-off in fog radio access networks under arbitrary demands. *IEEE Transactions on Wireless Communications* 18, 8 (Aug. 2019), 3871–3886.
7. Hamza, A.S. et al. A survey on inter-cell interference coordination techniques in OFDMA-based cellular networks. *IEEE Communications Surveys and Tutorials* 15, 4 (2013).
8. Khalid, M. et al. Communication through breath: Aerosol transmission. *IEEE Communications Magazine* 57, 2 (Feb. 2019), 33–39.
9. Letaief, B. et al. The roadmap to 6G: AI empowered wireless networks. *IEEE Communications Magazine* 57, 8 (2019), 84–90.
10. Mohammed, F. et al. Opportunities and challenges of using UAVs for Dubai smart city. In *Proceedings of the 6th International Conference on New Technologies, Mobility and Security (NTMS)*, Dubai, 2014.
11. Shihada, B. et al. Aqua-Fi: Delivering Internet underwater using wireless optical networks. *IEEE Communication Magazine: Design and implementation of Devices, Circuits, and Systems Series* 58, 5 (2020), 84–89.
12. Statista. Internet penetration rate in the Middle East compared to the global Internet penetration rate from 2009 to 2019. 2020; <https://www.statista.com/statistics/265171/comparison-of-global-and-middle-eastern-internet-penetration-rate/>
13. The Mohamed Bin Zayed International Robotics Challenge; <https://www.mbzirc.com>
14. United Nations Office for the Coordination of Humanitarian Affairs (OCHA). The future of technology in crisis response; <https://www.unocha.org/story/future-technology-crisis-response>.

Basem Shihada, CEMSE, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Tamer ElBatt, CSE, The American University in Cairo, New Cairo, Egypt.

Ahmed Eltawil, CEMSE, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia.

Mohammad Mansour, E&CE, American University of Beirut, Lebanon.

Essaid Sabir, NEST Research Group, ENSEM, Hassan II University of Casablanca, Casablanca, Morocco.

Slim Rkhis, CN&S, Sup'Com, University of Carthage, Tunis, Tunisia.

Sanaa Sharafeddine, CS, Lebanese American University, Beirut, Lebanon.

BY ASHRAF ABOULNAGA, AZZA ABOUZIED,
KARIMA ECHIHABI, AND MOURAD OUZZANI

Database Systems Research in the Arab World: A Tradition that Spans Decades

FROM HAMMURABI'S STONE tablets to papyrus rolls and leather-bound books, the Arab region has a rich history of recordkeeping and transactional systems that closely matches the evolution of data storage mediums. Even modern-day data management concepts like data provenance and lineage have historic roots in the Arab world; generations of scribes meticulously tracked Islamic prophetic narrations from one narrator to the next, forming lineage chains that originated from central Arabia.

Database systems research has been part of the academic culture in the Arab world since the 1970s. High-quality computer science and database education

was always available at several universities within the Arab region, such as Alexandria University in Egypt. Many students who went through these programs were drawn to database systems research and became globally prominent, such as Ramez Elmasri (professor at University of Texas, Arlington), Amr El Abbadi (professor at University of California, Santa Barbara), and Walid Aref (professor at Purdue University). Some have commented that it was easy to get into database research because it is a microcosm of computer science. The big tent of database research encompasses all: from systems to theory to languages and query optimization.

Today, many prominent database researchers are settling within the region and furthering its database research culture. For example, Ahmed Elmagarmid at the Qatar Computing Research Institute (QCRI) received the 2019 SIGMOD Contributions Award, Azza Abouzied at New York University Abu Dhabi (NYUAD) in the U.A.E. received the 2019 VLDB Test of Time Award, and of the six ACM Distinguished Members and Fellows in the region at the time of this writing, three are database systems researchers: Ashraf Aboulnaga, Ahmed Elmagarmid, and Mohamed Mokbel (all at QCRI). While the reasons for establishing their research groups within the region may vary, there are a few main influencing factors. First, there is a growing, collaborative nucleus of researchers spread throughout the region. Second, there is a thirst for establishing global research universities locally or in partnership with international universities. Third, there are funding opportunities and supportive environments for high-impact research. Finally and most importantly, there are many eager graduate students and young researchers interested in data research.

We describe the works of a few researchers to illustrate the diversity and richness of database systems research in the region. We begin our tour by exploring research on data preparation



and cleaning, followed by research on graph data management and advanced analytics. We conclude with research that aims to protect individuals' privacy as our society's reliance on data-driven systems continues to grow.

The research we present addresses problems of broad interest to the global database community and is not necessarily specific to the Arab world. Nevertheless, much of the work has regional relevance. For example, data preparation and cleaning are required for local datasets, and Arab societies are paying increasing attention to issues of fairness and data privacy. Like their predecessors, today's researchers in the Arab region also are training future generations of database systems researchers, who can use their knowledge to address pressing data management problems both regionally and globally.

Data Preparation and Cleaning

Data preparation, including chiefly data discovery, integration, and clean-

ing, consumes a disproportionate amount of time in data science tasks. In particular, data cleaning is hard to formalize and as a result, in practice, most data cleaning is carried out in an ad hoc and non-reproducible manner.

During the last decade, a team at QCRI has done influential work in this area. In particular, the team built NADEEF,⁷ an early precursor of several contemporary data cleaning systems that uses generic data quality rules that a given dataset must satisfy to identify errors and repair them. NADEEF was later extended¹⁶ to handle extremely large datasets by judiciously taking the data quality rules into a series of transformations that enable distributed computations and several optimizations, such as shared scans and specialized joins operators.

Another data preparation challenge relates to entity resolution, identifying records that refer to the same real-world entity. A critical challenge is that records come in different shapes and forms, which makes matching

them hard even for humans. DeepER⁹ was the first to use deep learning to solve this problem by capturing the semantics of records using distributed representations of words, also known as word embeddings. DeepER converts each record into a distributed representation (that is, a vector) using uni- and bi-directional recurrent neural networks (RNNs) with long-short-term-memory (LSTM) hidden units effectively capturing similarities between records.

Finally, the team built, in collaboration with MIT, Data Civilizer,⁸ an end-to-end system to support the entire life cycle of data preparation while tying it to downstream analytic applications, especially machine learning applications. The premise is that data in any organization is scattered among a multitude of databases, data lakes, and so on, and there is a need for a system to find the data of interest, integrate it, and clean it in a scalable fashion. Data Civilizer provides several modules to address these different steps.

High-quality computer science and database education was always available at several universities within the Arab region.

Graph Data Management

Graph data is ubiquitous, from the Web to advertising to biology, and the development of algorithms and systems for graph management and analytics has spawned a large amount of research in the region.

Panos Kalnis and his group at King Abdullah University of Science and Technology (KAUST) in Saudi Arabia have built several graph systems. They recently proposed to use sparse matrix algebra as a design paradigm for graph query engines. They built MAGiQ,¹⁴ a system that represents an RDF graph as a sparse matrix and uses matrix algebra to execute queries on graphs with hundreds of billions of edges, scaling to thousands of compute nodes.

Arabesque,¹⁸ developed by a team at QCRI, was perhaps the first system to make it possible to carry out large-scale graph data mining tasks, like frequent subgraph mining, in a principled fashion. Arabesque reconceptualized graph analysis by introducing the paradigm of “think like an embedding” instead of “think like a vertex,” which had been the standard approach for graph analysis. This new way of perceiving graph analysis made it possible to create a succinct API for many graph mining tasks and a scalable implementation of this API in a cluster setting.

LiveGraph,¹⁹ developed by QCRI and Tsinghua University, is a system designed to simultaneously support transactions and complex analytics on graphs. The key innovation in LiveGraph is a novel data structure that stores edges contiguously and supports efficient edge scans and multi-version concurrency control. This allows high-speed, concurrent processing of transactional and analytical workloads on the primary graph store without the need for expensive extract-transform-load processes.

Mohammad Hammoud and his team at Carnegie Mellon University in Qatar (CMU Qatar) have developed an open source, cloud-based distributed system for graph analytics called LA3.¹ Like KAUST’s MAGiQ system, LA3 uses a highly optimized linear algebra-based execution engine. It provides a familiar vertex-based programming model and was later extended with a

novel architecture-aware parallelism model for high-performance computing platforms, substantially outperforming the state of the art.

Kamel Boukhalfa and his group at the University of Science and Technology Houari Boumediene in Algeria have worked on the related area of databases and cloud computing. They recently proposed an approach to optimize data placement in a hybrid storage system.³ The approach considers several sub-costs, such as occupancy cost, durability cost, and migration cost.

Karam Gouda and his students at Benha University in Egypt have made several contributions to graph edit-based similarity search queries. Due to the hardness of computing sub-graph isomorphism and graph edit distance, these queries are often executed using a filter-and-verify approach: an efficient approximate algorithm is first used to identify a set of candidate results, then an expensive verification process on the candidates is applied to compute the final results. The group proposed an efficient verification method that can work as a stand-alone graph edit-based similarity search and outperforms the state of the art by over two orders of magnitude.¹³

Prescriptive Analytics

Azza Abouzied at NYUAD is trying to shift the data analytics paradigm from descriptive and predictive to prescriptive. Currently, database systems do not natively support the many data processing needs of data-driven decision making, leaving experts to develop their own custom, ad hoc application-level solutions that are difficult to scale and may produce sub-optimal results. While many systems provide support for scalable descriptive analytics (like statistics and summaries of the raw data) and even some predictive analytics (such as forecasts), there is little support for prescriptive analytics, which searches for the best course of action given the available data. As we move from “what is the data?” to “what to do with it?,” Abouzied and her colleagues at the University of Massachusetts Amherst are augmenting database systems with efficient computational problem-solving capabilities^{4,5} that take into consideration the inherent uncertainty of

data and models.⁶ In particular, they are integrating state-of-the-art solvers within the DBMS to scalably solve stochastic constrained optimization problems with tight approximation guarantees. Abouzied's research has won several awards, such as the CACM 2019 Research Highlight Award, and she is applying it to current problems facing the region: her team is building a tool that advises policymakers on how to cost-effectively control epidemics like the current COVID-19 pandemic with data-driven prescriptive analytics.

Scalable, Accurate Analytics

Karima Echihabi at Mohammed VI Polytechnic University in Morocco is tackling fundamental problems to facilitate scalable and accurate analytics, focusing on supporting efficient similarity search for very large collections of high-dimensional vectors. One particular work demonstrates that it is possible to design efficient high-dimensional vector similarity search algorithms with theoretical guarantees on the quality of the answers,¹¹ and thus offers a more promising alternative than the current state of the art. She has conducted the two most extensive experimental evaluations in the area of similarity search for data series and generic high-dimensional vectors, offering novel insights into this challenging problem and identifying new promising research directions.^{10,11} Also, her work on progressive query answering, in collaboration with colleagues from Université de Paris and Inria, has led to techniques that support interactive exploration and fast decision-making on massive data series collections.¹²

Top-k Information Retrieval Queries

Information retrieval is an important data management area, with contributions from several research groups in the Arab region. Shady Elbassouni at the American University of Beirut in Lebanon and his collaborators have worked on quantifying and addressing fairness in online job search platforms. They have built various frameworks to reveal and compare the fairness of different jobs at different locations for different demographic

groups. With the aid of top-k style query-processing algorithms, they were also able to retrieve the jobs, locations, or groups for which a job search platform is the least or most fair.²

Hammoud's group at CMU Qatar with Tamer Elsayed from Qatar University also worked on top-k queries in information retrieval systems. They recently conducted the first extensive comparison of 10 effective information retrieval strategies under five representative ranking models. Based on this comparison, they proposed LazyBM,¹⁵ a simple query evaluation strategy that consistently and robustly outperforms all considered strategies.

Data Privacy

The final area we cover is data privacy. A team of researchers at QCRI, in collaboration with Yin "David" Yang at Hamad Bin Khalifa University in Qatar, is investigating local differential privacy, which allows data to be collected from users while providing strong privacy protection, even when the data collector cannot be trusted. Some of this team's recent work focuses on protecting sensitive relationship data in decentralized social networks, in which each user only has a view of their immediate neighborhood. For example, the contact lists in a group of users' phones form a decentralized social network. The team designed a suite of techniques to collect local views from users with local differential privacy. These techniques can be used for various graph analysis tasks, including link prediction and computing subgraph statistics.¹⁷

Conclusion

In this article, we offered a bird's-eye view of database systems research in the Arab world, spanning two continents from the Atlantic Ocean in the west to the Gulf in the east. We highlighted hubs of excellence, committed to tackling key research challenges with regional and global impact, training and empowering the next generation of researchers, and supporting technology transfer that can propel their societies forward. 

References

1. Ahmad, M.Y., Khattab, O., Malik, A., Musleh, A., Hammoud, M., Kutlu, M., Shehata, M. and Elsayed, T. LA3: A scalable link- and locality-aware linear

- algebra-based graph analytics system. In *Proceedings of VLDB Endow.* 11, 8 (2018).
2. Amer-Yahia, S., Elbassouni, S., Ghizzawi, A., Borromeo, R.M., Hoareau, E., and Mulhem, P. Fairness in online jobs: A case study on TaskRabbit and Google. In *Proc. Int. Conf. on Extending Database Technology*, 2020.
3. Boukhelef, D., Boukhozba, J., Boukhalifa, K., Ouarnoughi, H., and Lemarchand, L. Optimizing the cost of DBaaS object placement in hybrid storage systems. *Future Gen. Comput. Sys.* 93, 2019.
4. Brucato, M., Abouzied, A., and Meliou, A. Package queries: Efficient and scalable computation of high-order constraints. *The VLDB J.* 27, 5 (2018).
5. Brucato, M., Abouzied, A., and Meliou, A. Scalable computation of high-order optimization queries. *Commun. ACM* 62, 2 (2019).
6. Brucato, M., Yadav, N., Abouzied, A., Haas, P.J., and Meliou, A. Stochastic package queries in probabilistic databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2020.
7. Dallachiesa, M., Ebaïd, A., Eldawy, A., Elmagarmid, A.K., Ilyas, I.F., Ouzzani, M., and Tang, N. NADEEF: a commodity data cleaning system. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2013.
8. Deng, D., Fernandez, R.C., Abedjan, Z., Wang, S., Stonebraker, M., Elmagarmid, A.K., Ilyas, I.F., Madden, S., Ouzzani, M., and Tang, N. The Data Civilizer system. In *Proc. Conf. on Innovative Data Systems Research*, 2017.
9. Ebraheem, M., Thirumuruganathan, S., Joty, S., Ouzzani, M., and Tang, N. Distributed representations of tuples for entity resolution. *Proc. VLDB Endow.* 11, 11 (2018).
10. Echihabi, K., Zoumpatianos, K., Palpanas, T., and Benbrahim, H. The Lernaean hydra of data series similarity search: An experimental evaluation of the state of the art. *Proc. VLDB Endow.* 12, 2 (2018).
11. Echihabi, K., Zoumpatianos, K., Palpanas, T., and Benbrahim, H. Return of the Lernaean hydra: Experimental evaluation of data series approximate similarity search. *Proc. VLDB Endow.* 13, 3 (2019).
12. Gogolou, A., Tsalidas, T., Echihabi, K., Bezerianos, A., and Palpanas, T. Data series progressive similarity search with probabilistic quality guarantees. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2020.
13. Gouda, K., and Hassaan, M. CSI_GED: An efficient approach for graph edit similarity computation. In *Proc. IEEE Int. Conf. on Data Engineering*, 2016.
14. Jamour, F.T., Abdelaziz, I., Chen, Y., and Kalnis, P. Matrix algebra framework for portable, scalable and efficient query engines for RDF graphs. In *Proc. EuroSys Conf.*, 2019.
15. Khattab, O., Hammoud, M., and Elsayed, T. Finding the best of both worlds: Faster and more robust top-k document retrieval. In *Proc. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2020.
16. Khayyat, Z., Ilyas, I.F., Jindal, A., Madden, S., Ouzzani, M., Quiane-Ruiz, J.-A., Papotti, P., Tang, N., and Yin, S. BigDancing: a system for big data cleansing. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2015.
17. Sun, H., Xiao, X., Khalil, I., Yang, Y., Qin, Z., Wang, W.H., and Yu, T. Analyzing subgraph statistics from extended local views with decentralized differential privacy. In *Proc. ACM SIGSAC Conf. on Computer and Communications Security*, 2019.
18. Teixeira, C.H., Fonseca, A.J., Serafini, M., Siganos, G., Zaki, M.J., and Aboulmaga, A. Arabesque: a system for distributed graph mining. In *Proc. Symp. on Operating Systems Principles*, 2015.
19. Zhu, X., Serafini, M., Ma, X., Aboulmaga, A., Chen, W., and Feng, G. LiveGraph: A transactional graph storage system with purely sequential adjacency list scans. In *Proc. VLDB Endow.* 13, 7(2020).

Ashraf Aboulmaga is a Senior Research Director at the Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar.

Azza Abouzied is an assistant professor of computer science at New York University, Abu Dhabi, U.A.E.

Karima Echihabi is an assistant professor of computer science at Mohammed VI Polytechnic University, Morocco.

Mourad Ouzzani is a Principal Scientist at the Qatar Computing Research Institute, Hamad Bin Khalifa University, Qatar.

Copyright held by authors/owners.
Publication rights licensed to ACM.

BY AHMED ALI, SHAMMUR CHOWDHURY,
MOHAMED AFIFY, WASSIM EL-HAJJ, HAZEM HAJJ,
MOURAD ABBAS, AMIR HUSSEIN, NADA GHNEIM,
MOHAMMAD ABUSHARIAH, AND ASSAL ALQUDAH

Connecting Arabs: Bridging the Gap in Dialectal Speech Recognition

AUTOMATIC SPEECH RECOGNITION refers to the process through which speech is converted into text. Over the decades, automatic speech recognition has achieved many milestones, thanks to advances in machine learning and low-cost computer hardware. As a result, the best systems for English have achieved a single-digit word error rate (WER) and, in some conversational tasks, performance is comparable to human transcribers. This led researchers to debate

whether the machine has reached human parity in speech recognition.^{9,16}

Unlike English, speech recognition in Arabic faces many challenges, even with such advanced techniques. Arabic poses a set of unique challenges due to its rich dialectal variety, with modern standard Arabic (MSA) being the only standardized dialect.⁴

MSA is syntactically, morphologically, and phonologically grounded on classical Arabic, the language of the Qur'an (Islam's Holy Book). Lexically, however, it is much more modern.⁸ MSA is taught in schools across the Arab region and is the main language in news broadcasts, parliaments, and formal speech. This is one of the main reasons why MSA has been the main choice for speech and language technology for the last two decades. The current WER for MSA automatic speech recognition (ASR) is about 13%,^{2,9} and is worse for dialectal ASR, where the WER averages 30%.^{1,3}

Remarkably, the 400 million Arabic native speakers (estimated in 2020) use Dialectal Arabic (DA) as their means of communication in day-to-day speech. MSA is not the native language of any Arab. Dialects used to be primarily spoken, not written. However, this has changed with the rise of Web 2.0, when DA became a written, as well as a spoken, language.

An objective comparison of the varieties of Arabic dialects could potentially lead to the conclusion that Arabic dialects are historically related, and that they are not mutually intelligible languages like English and Dutch. Normal vernacular can be difficult to understand across different Arabic dialects. The tomato example in Figure 1 shows lexical variation across multiple Arabic countries.⁵ How different is tomato in Arabic compared to English? In English, there is one lexical form for tomato and two phonological variations, with a 16.7% difference between **to-MAY-to** and **to-MAH-to**, while DA has 10 lexical variations with 67% average-character difference and 15 phonological varia-



tions with 87% average-phoneme difference. Despite the fact that there has been a great deal of speech recognition research in MSA, there is limited effort toward building a platform with standard lexicon and training data to benchmark results and to advance the state of the art in Dialectal ASR.

Dialectal spoken Arabic poses three main challenges: *lack of resources*, *lack of standard orthographic rules*, and *lack of definition* (that is, Arabic is a language with many dialects or sets of languages).

Spoken Arabic Dialect Identification

From the speech perspective, dialectal

variation increases the level of ambiguity, due to the addition of different linguistic and acoustic representations.

Arabic Dialectal Corpora: There have been numerous efforts to produce *spoken Arabic data set resources*. The CallHome task within the 1996–1997 NIST benchmark evaluations framework¹⁵ reported the first transcribed Arabic dialect dataset. In 2003 and 2004, NIST evaluations provided more DA data, mainly in the Egyptian and Levantine dialects, as part of language recognition evaluation. Data from the Iraqi dialect was first obtained as resources for two main research programs: global

autonomous language exploitation (GALE),¹⁴ a U.S. Defense Department Defense Advanced Research Projects Agency (DARPA) program carried out between 2006 and 2009, and the spoken-language communication and translation system for tactical use (TRANSTAC) program, aimed to help the U.S. soldier communicate with non-English speakers using a portable bidirectional translator. These datasets exposed the research community to the challenges in spoken DA.

One of the main challenges of processing dialectal speech is to first identify the dialect of the spoken content. Arabic has more than 30 dialects that can be categorized geographically, socially, or phonologically. However, obtaining complete coverage of major dialects is still challenging.

For the Arabic dialect identification (ADI) task, the multi-genre broadcast (MGB-3) challenge¹ in ASRU-2017 provided a dataset, ADI-5, containing four regional Arabic dialects and MSA, with 53 hours of speech as training data. As a continuation of enriching the DA datasets, with fine-grained analysis of dialectal Arabic speech in the MGB-5 challenge,³ the ADI-17 dataset was released. This includes about 3,000 hours of dialectal Arabic crawled from YouTube covering 17 Arabic-speaking



Unlike English, speech recognition in Arabic faces many challenges due to its rich dialectal variety, with modern standard Arabic (MSA) being the only standardized dialect.

countries. Furthermore, another 58 hours of speech, manually annotated, was provided as development and test sets of ADI-17. This is currently the largest available spoken dialectal Arabic dataset for ADI. Figure 2 shows mapping between ADI-17 and ADI-5. Further data on available speech for spoken dialectal Arabic is provided in the accompanying table.

Dialectal systems: To design the **dialect identification system**, the current state of the art adopts many approaches developed for speaker and language recognition.

The task is explored using different supervised and semi-supervised architectures covering diverse topics ranging from domain adaptation to end-to-end learning.

Attention also was given to linguistic feature extraction. For acoustic representation, techniques ranged from simple frame-level spectral features to i-/x-vectors latent representations. Many researchers explored different data augmentation techniques, such as speed and volume perturbation, to increase the diversity and amount of training data. Others exploited approaches such as time-scale modification for balancing the low-resource dialect. The most popular architecture, based on the latest MGB-5 ADI challenge, is convolutional neural network. A high performance of around 95% accuracy is seen in the broadcast domain (the best result from the MGB-5 challenge) using this CNN architecture.

The accuracy of such a model shows closeness between some dialects with shared features while also highlighting the discriminating

characteristics between different dialectal forms of Arabic. Figure 3 shows the confusion matrix for the 17 Arabic dialects. However, this performance can differ across different channels; for example, the performance of modeling narrowband telephony data is less accurate than the broadband broadcast domain. With respect to deploying ADI in ASR, it is argued both ways: building a different ASR system for each dialect versus combining all dialects in a single unified system.⁷

Arabic Speech Recognition by Human and Machine

In English, **enough** is the correct spelling, and **enuf** is a wrong one, although its pronunciation is close enough. In Arabic, MSA transcription has twice as many errors as in English. This is mainly due to the lack of diacritization, which causes problems particularly in determining the location of the vowels. The high degree of complexity in Arabic morphology causes a high degree of affixation. English reports a 5.8% Human error rate,¹⁶ while broadcast news in Arabic has 10%.⁹

The phrase “as I told you” can have more than 20 lexical representations in DA. In the DA speech-recognition challenges, authors reported that four native speakers had 15% inter-annotation disagreement for the Egyptian dialect¹ and 40% inter-annotation disagreement for Moroccan dialects.³ This is the upper-bound WER for a perfect ASR system, assuming WER is an appropriate metric to evaluate dialectal ASR.

The speech team at the Qatar Computing Research Institute (QCRI)



and the Massachusetts Institute of Technology Computer Science and Artificial Intelligence Lab (MIT CSAIL) have built several Arabic ASR systems. They explored grapheme and phoneme representations for acoustics units and hidden Markov models with time-delay neural network for acoustic modeling; for language units, they investigated word, character, and word-pieces; and for language modeling, N-Gram and recurrent neural networks have been studied.^{10,11,13}

Recently, researchers from Kanari AI, QCRI and John Hopkins University developed an end-to-end multi-dialect Arabic ASR system using transformer architectures. Their research led to 12.5%, 27.5%, 33.8% WER; a new performance milestone for the broadcast news, the Egyptian, and the Moroccan ASR challenges respectively. It was noticeable that the mistakes produced by the end-to-end transformer showed high similarity with the expert linguist transcription. However, their results suggest that human performance in the Arabic language is still considerably better than the machine, with an absolute WER gap of 3.6% on average.⁹

The WER for MSA achieved less than 13% in the MGB-2. This is now trusted by multiple users, such as the Al Jazeera network, BBC media-monitoring, and many government entities, thanks to the 1,200 transcribed hours and the 130 million words shared by the MGB-2 challenge, which is the case of broadcast news. Nevertheless, the wild DA is still lagging behind, with an average of 28% WER in MGB-3 and 34% in MGB-5.

Code-switching is one of the main challenges in spoken DA. In an eight-hour corpus collected over two days of meetings of the United Nations Economic and Social Commission for West Asia (ESCWA) in 2019, it was observed that more than 2.5 hours of the collected speech demonstrated intrasentential code-switching, where the alternation between Arabic and English is happening within the same sentence. In some cases, as with Algerian, Tunisian, and Moroccan native-speakers, the switch is between Arabic and French. In cases like this, there is an urge for a bilingual ASR, rather than just a robust multilingual ASR for the dialectal Arabic content.

Most notable multi-dialectal Arabic speech datasets. With * representing the datasets freely available.

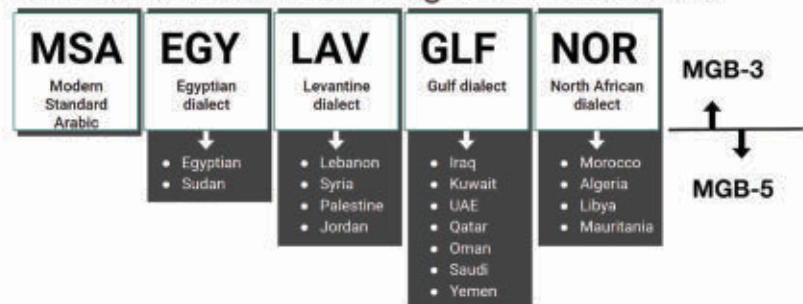
Datasets	Channels	Dialect Labels	Duration
Regional/Country Level Dialect Variation			
ADI-5*	Broadcast News	5 (Regional)	74h
VarDial2018* (only test set is available)	Multimedia (YouTube)	5 (Regional)	26h
GALE Phase 2 Arabic Broadcast Conversation Speech	Broadcast News	2 (MSA vs Dialect)	251h
Multi-Language Conversational (Telephone Speech 2011)	Telephone	4 (Regional)	117h
NIST LRE 2017 (most recent from the series)	Telephone	4 (Regional)	-
ADI-17*	Multimedia (YouTube)	17 (Arabic countries)	3091h
Within Dialect Variation			
TARIC	Recorded conversation	1 (Tunisian)	20h
KALAM'DZ	Multimedia (YouTube, Online Radio and TV)	8 Algerian dialectal variation (Hilali-Saharan, Hilali-Tellian, High-plains, Ma'qilian, Sulaymite, Algiers Blanks, Sahel-Tell, and Pre-HilaliTunisian)	104h
AMCASC	Telephone	5 Algerian cities (Constantine, Oran, Algiers, Kabyle, and Saharian)	88h

Figure 1. Sample lexical variations for a single word in dialectal Arabic across countries in the Arab region.



Figure 2. Mapping between ADI-5 and ADI-17.

• Previous datasets has 5 regional dialect class



-> Not enough to cover Arab world

Figure 3. Confusion matrix for the ADI-17 challenge, with an overall accuracy: 82%.

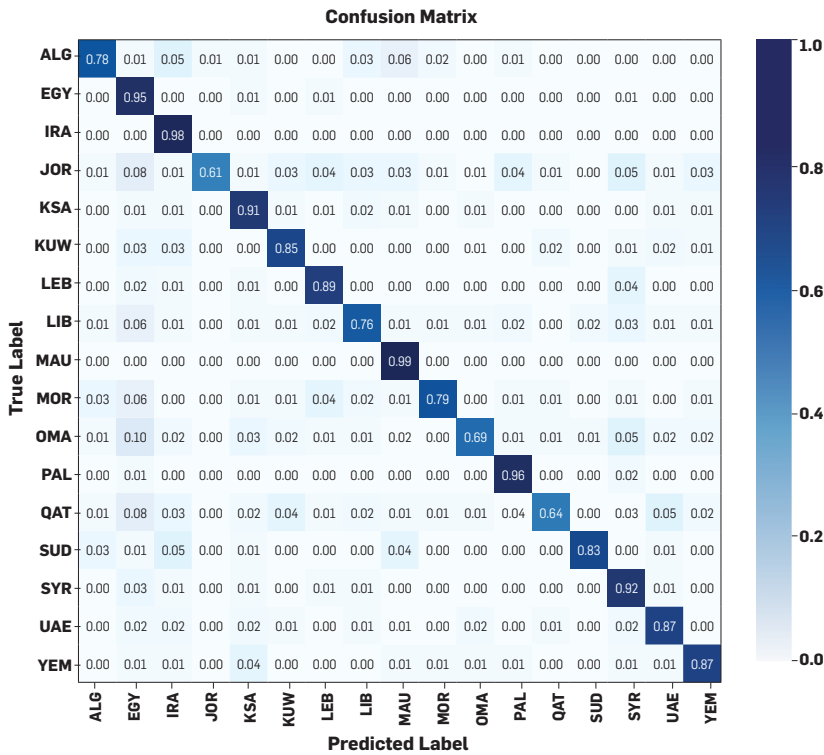


Figure 4. Example shows the integration of speech recognition, dialect identification, MT and NLP stacks happening in real time.

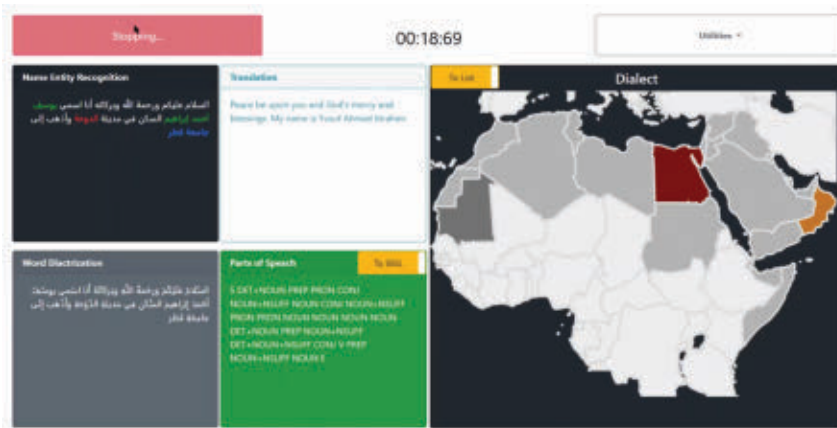


Figure 5. The first ArabicSpeech meeting in 2019 at QCRI.



Connecting dialectal speech with NLP modules: Figure 4 shows an application for dialectal speech processing pipeline; FarSpeech⁶ is a system that integrates multiple language technologies starting from the voice. Initially, the application will detect which Arabic country the speaker is from using acoustic and lexical features, and based on the chosen dialect, it will choose the ASR system to convert speech into text. Secondly, this application will apply the natural language processing (NLP) stack, as part of speech tagging, named entity recognition, and vowelization restoration. Finally, the recognized text is translated into English.

The Arabic Speech Community
 Recently, there has been an effort to connect the research community working on Arabic speech. *ArabicSpeech*,^a founded in 2018, is an emerging research community for the benefit of Arabic speech science and speech technology. ArabicSpeech is concerned with technologies in speech and language processing focusing on both standard and dialectal Arabic language. ArabicSpeech aims to help build innovation and technology capacities for the Arabic language. It focuses on tackling large-scale computing challenges that address real priorities impacting people’s lives. The community aims to push the boundaries of Arabic speech technologies.


Before 2018: Earlier efforts for Arabic speech recognition were uncoordinated. In particular, there were three important projects:

- ▶ The 2002 Johns Hopkins Summer Workshop¹² focused on Arabic ASR.
- ▶ The GALE project was funded by DARPA to produce a system able to automatically transcribe, translate, and summarize multilingual newscasts. Arabic speech recognition was one of the core technologies of the GALE project, which was mainly concerned with Arabic broadcast news and broadcast conversation.
- ▶ The MGB challenges are evaluations of speech recognition, speaker diarization, dialect identification, and lightly supervised alignment using TV

a <https://arabicspeech.org/>



approaches to improve speech recognition systems using unlabeled data.

In summary, there is a need to build a sizeable dialectal speech corpus in the wild, and to investigate techniques for dialectal speech processing. 

References

1. Ali, A. et al. Speech recognition challenge in the wild: Arabic MGB-3. *IEEE Automatic Speech Recognition and Understanding Workshop*, (2017), 316–322.
2. Ali, A. et al. The MGB-2 challenge: Arabic multi-dialect broadcast media recognition. *IEEE Spoken Language Technology Workshop* (2016), 279–284.
3. Ali, A. et al. The MGB-5 challenge: Recognition and dialect identification of dialectal Arabic speech. *IEEE Automatic Speech Recognition and Understanding Workshop* (2019), 1026–1033.
4. Badawi, E.S. et al. *Modern Written Arabic: A Comprehensive Grammar*. Routledge, 2013.
5. Bouamor, H. et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the 11th Intern. Conf. Language Resources and Evaluation* (2018).
6. Eldesouki, M. et al. *FarSpeech: Arabic Natural Language Processing for Live Arabic Speech*. INTERSPEECH (2019), 2372–2373.
7. Elfeky, M.G. et al. Multi-dialectal languages effect on speech recognition: Too much choice can hurt. *Procedia Computer Science* 128, (2018), 1–8.
8. Habash, N.Y. 2010. Introduction to Arabic natural language processing. *Synthesis Lectures on Human Language Technologies* 3, 1 (2010), 1–187.
9. Hussain, A. et al. Arabic Speech Recognition by End-to-End, Modular Systems and Human; arXiv (Jan. 2020).
10. Khurana, S. et al. DARTS: Dialectal Arabic Transcription System. arXiv:1909.12163 (Sep. 2019).
11. Khurana, S. and Ali, A. QCRI advanced transcription system (QATS) for the Arabic Multi-Dialect Broadcast media recognition: MGB-2 challenge. 2016 IEEE Spoken Language Technology Workshop (SLT) (San Diego, CA, Dec. 2016), 292–298.
12. Kirchhoff, K. et al. Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop. In *Proceeding of the 2003 IEEE Intern. Conf. Acoustics, Speech, and Signal Processing*.
13. Najafian, M. et al. Automatic speech recognition of Arabic multi-genre broadcast media. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (2017), 353–359.
14. Olive, J. et al. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media, 2011.
15. Pallett, D.S. A look at NIST's benchmark ASR tests: past, present, and future. 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721) (2003), 483–488.
16. Xiong, W. et al. 2016. Achieving human parity in conversational speech recognition. arXiv preprint arXiv:1610.05256. (2016).

Ahmed Ali, Qatar Computing Research Institute, HBKU, Qatar.

Shammur Chowdhury, Qatar Computing Research Institute, HBKU, Qatar.

Mohamed Afify, Microsoft Advanced Technology Lab, Egypt.

Wassim El-Hajj, American University of Beirut, Lebanon.

Hazem Hajj, American University of Beirut, Lebanon.

Mourad Abbas, Centre de Recherche Scientifique et Technique pour Le Développement de la Langue Arabe, Algeria.

Amir Hussein, Kanari AI, UAE.

Nada Ghneim, Arab International University, Syria.

Mohammad A.M. Abushariah, King Abdullah II School of Information Technology, The University of Jordan, Jordan.

Assal Alqudah, Taibah University, Saudi Arabia.

Copyright held by authors/owners.

recordings and YouTube data. MGB-2 focused on broadcast news MSA data, while MGB-3 and MGB-5 focused on dialectal Arabic speech challenges in the wild using YouTube multi-genre speech content.

These have been great initiatives, with good research impact, but not sustainable as they rely on timely funded projects, such as DARPA, SUMMA (an EU-funded project),^b or a co-located challenge, as with one of the speech conferences.

ArabicSpeech: In 2018, ArabicSpeech was founded by an advisory board with a mixture of researchers from academia and industry: QCRI, Johns Hopkins University, University of Edinburgh, Google, and Microsoft. This mix of academia and industry reflects the challenges and activities organized by ArabicSpeech. The community organizes an annual workshop to discuss ongoing projects related to Arabic speech. We created a special interest group in the International Speech Communication Association (ISCA). The platform now gives a one-stop location for young researchers and practitioners interested in Arabic speech. Since founding it, most of the speech teams in major companies have become interested in helping the community, with more than 300 researchers worldwide joining the ArabicSpeech community. Today, more than 4,000 hours have been shared on the ArabicSpeech resource platform. Figure 5 shows the first meeting of ArabicSpeech, organized in 2019.

At the time of writing this article,

^b <http://summa-project.eu/>

we can recognize growing labs working on Arabic speech recognition-related challenges in the Arab world, including QCRI in Qatar; Kanari AI in Qatar and UAE; Microsoft Advanced Technology Lab, RDI, and Cairo University in Egypt; American University of Beirut in Lebanon; Ecole Normale de Bouzareah and CRSTDLA (The Scientific and Technical Research Center for the Development of the Arabic Language) in Algeria; HIAST and Damascus University in Syria, and King Abdullah II School of Information Technology, The University of Jordan.

What's Next?

There are many gaps remaining in making Arabic ASR practically useful, such as home assistance devices that can understand and speak Arabic dialects. State-of-the-art ASR systems have been trained on tens of thousands of hours transcribed verbatim, or at least semi-supervised, by choosing the most likely word sequence for each utterance. Unfortunately, Arabic dialects still need large corpora for every dialect. Most of the state-of-the-market systems are using broadcast speech data such as GALE, MGB, or the SUMMA corpus, which are mainly formal speech. In addition to the development of large-scale datasets, there is a need for a holistic approach to combine different perspectives of multi-dialect Arabic speech processing systems capable of recognizing dialect, converting it to text via ASR, and finally communicating it back in dialectal Arabic—dialectal text to speech. It is arguably convincing that self-training and unsupervised pretraining have emerged as effective

Article development led by [acmqueue](https://queue.acm.org)
queue.acm.org

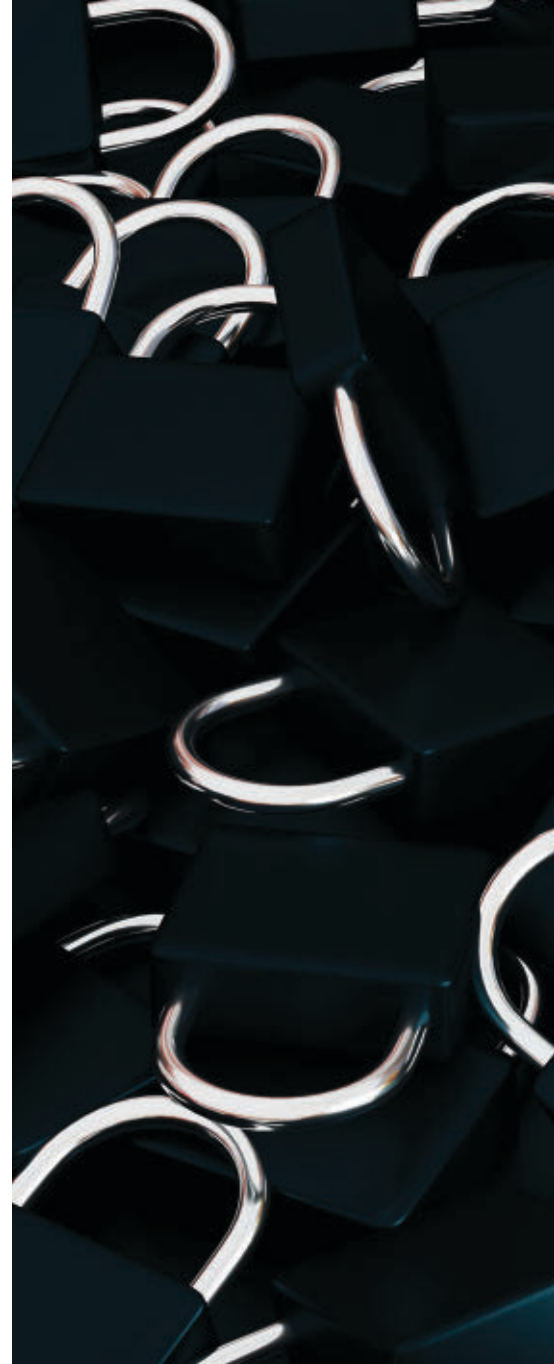
The 24-year-old security model has found a second wind.

BY DAVID CRAWSHAW

Everything VPN Is New Again

THE VIRTUAL PRIVATE network (VPN) is 24 years old. The concept—cryptographically secure tunnels used as virtual wires for networking—was created for a radically different Internet from the one we know today. As the Internet grew and changed, so did VPN users and applications. The VPN had an awkward adolescence in the Internet of the 2000s, interacting poorly with other widely popular abstractions such as multiuser operating systems. In the past decade the Internet has changed again, and this new Internet offers new uses for VPNs. The development of a radically new protocol, WireGuard, provides a technology on which to build these new VPNs.

This article is a narrative history of the VPN. All narratives necessarily generalize and cannot capture every nuance, but it is a good-faith effort to (critically) celebrate some of the recent technical history of networking and to capture the mood and attitudes of software engineers and network administrators toward the VPN.



The First Age: Fiefdoms and Leased Lines

Before the Internet there were networks: corporate networks, university networks, government networks. These networks were made of relatively expensive computers, had relatively few trusted people (at least by the standards of today's multibillion-person Internet), were managed by full-time network administrators, and were geographically clustered into buildings or campuses.

When an organization was split across more than one site it connected its networks with a leased line. In the 1970s this was dedicated unswitched copper wiring provided by a phone network to run a proprietary



IMAGE BY DEN RISE

protocol such as DECnet.³ Leasing a physical wire across hundreds or thousands of miles was not cheap. As phone networks became more sophisticated the leased line evolved into frame relays and connections to an ATM (asynchronous transfer mode) network. These networks reduced the cost of leased lines from the astronomical prices affordable by only the largest enterprises to merely very high prices accessible to a few more big companies.

The security model was physical and contractual. On-site networks were kept safe because the wall jacks into the network were guarded by guards and access badges. The leased lines were similarly guarded, so went the

theory, by the phone company. Large contracts certainly felt safe.

Through the 1980s and 1990s the Internet was busy being built, under many names and in various places, by organizations interconnecting their networks using leased lines and by ISPs offering relatively cheap dial-up access to one of these peering networks. Many smaller organizations spread across multiple sites could not afford a leased line but could afford to have each of their sites connected to a local ISP.

This raised an interesting possibility: Could a smaller organization get the benefit of a leased line connecting its sites over the internet? The VPN was born.

PPTP. Several projects in the early 1990s worked on IP-layer security. The first one that could be called a VPN was swIPE.⁶ The draft standard dates it to 1993. In swIPE, IP datagrams are encapsulated for encryption and then transmitted over another network, so you can make the claim that this is a VPN. It was never widely deployed.

The first unambiguous VPN was Point-to-Point Tunneling Protocol (PPTP). It was created in 1996 and standardized in RFC 2637 in 1999.⁵

Fittingly, PPTP was the product of a company that, at the time, produced networking software used by smaller businesses: Microsoft. (Microsoft did not have a stellar reputation as a network company in the 1990s, though to-

day it runs, in terms of revenue, the second-largest cloud provider.)

PPTP worked, in its way. It played a great game of pretend, packaging up Point-to-Point Protocol (PPP), encrypting the stream, and running it over a TCP (transmission control protocol) socket. Working as a virtual PPP, PPTP was able to encapsulate several network protocols, including a popular alternative to TCP/IP at the time among smaller organizations: Internet Packet Exchange (IPX), the protocol used by Novell's NetWare.

The cryptographic algorithms employed by PPTP—RC4 (Rivest Cipher 4) and DES (Data Encryption Standard)—are long obsolete, but even when these algorithms were considered adequate, several flaws in PPTP's implementation created security vulnerabilities. Such vulnerabilities would become an ongoing thorn in the side of the VPN.

IPSec. The 1990s were hectic. In the miasma of 1993 (top chart song that year: "Achy Breaky Heart"), the Internet Engineering Task Force (IETF), an open-standards organization fresh from victory standardizing TCP/IP version 4, formed an IPSec (IP security) committee. IETF's goal was to bring security to IP. This was a much broader focus than creating what we know of today as a VPN, and the result was a standard that does a lot of things.

The first-draft RFCs (requests for comments) from the IETF for an all-purpose secure IP encapsulation standard were published in 1995. A working prototype came after PPTP, and the specification was published in RFC 2401 in 1998⁷ with implementations starting to appear shortly thereafter.

A historical aside: the first-draft specifications of IPSec predate PPTP, and the first prototypes postdate PPTP, which raises the question of whether there was any connection between these two projects. It seems unlikely; instead this appears to be a case of history happening all at once. Indeed, two other VPN protocols were developed between PPTP's creation in 1996 and standardization in 1999. L2F (Layer 2 Forwarding) was a Cisco PPP-over-IP protocol developed in 1997 and standardized in RFC 2341.¹¹ L2TP (Layer 2 Tunneling Protocol) is another

protocol that borrows from both L2F and PPTP, though it was not standardized until later. Even more confusingly, the first RFC to use the term *VPN* is RFC 2194, published in 1997.¹ At this point the VPN had existed for only a year, but this RFC mentions three different protocols: PPTP, L2F, and L2TP (but not IPSec).

IPSec does everything, and everything it does is configurable. It has two modes of operation—tunnel and transport—multiple cryptographic suites, and multiple independent implementations. The result is an all-purpose toolkit for constructing networks that is still in use today. Like all things that try to be everything to everyone, however, for many of us the prospect of setting it up (and, more importantly, maintaining it) is daunting.

The Second Age: Satellite Offices and Consumer Privacy

By the early 2000s in the U.S., it was possible for almost all desktops and laptops to reach the Internet, though many remained disconnected for reasons of policy, price sensitivity, or lack of adequate networking software. (Often this lack of software meant not that it was impossible to route a local network onto the Internet, but that doing so required a great deal of manual intervention by an expert, and experts were in short supply.)

More people wanted to connect offices together; small businesses wanted their employees to reach their exchange servers from satellite offices; the early (and at the time rare) remote workers wanted the same experience at home or on the road that they got in the office. The VPN had to become easier to configure and integrate with a user-friendly authentication scheme.

SSL/TLS. In a separate development, Web browsers had developed a robust means of encrypting traffic: SSL (Secure Sockets Layer), later TLS (Transport Layer Security), which involved distributing a trusted set of root certificates to clients used to identify servers. With Web browsers finding their way onto all computers, reusing these certificates for VPNs provided an easy way to deploy VPN gateways that remote users and satellite offices could connect to in a hub-and-spoke arrangement.

One notable open-source VPN product built on these principles is OpenVPN, which lets users authenticate with a username/password and connect to a VPN gateway that is authenticated with a TLS certificate. This project is active and forms the foundation of many corporate and consumer VPNs today.

Consumer VPNs. As more consumer activity moved onto the Internet, traditional businesses that relied on restricting consumer access across geographies to maximize revenue began to introduce services locked to regions, mechanically enforced by looking at the user's IP address. Simultaneously, a new industry of targeted advertising was developed that tracks users by using their IP addresses to determine interests and spending habits.

In response to these industries, consumer VPN products were developed and became moderately popular. These are very different from the corporate VPN, whose objective is to move a packet between two trusted endpoints without revealing it to intermediate carriers. The consumer VPN, on the other hand, hides traffic from the consumer to the VPN gateway, and then as a proxy for the consumer sends the packet to a public server. This hides the consumer's IP address, confounding these new industries.

This new product introduces an interesting terminology challenge when talking about VPNs. The technology is the same—an encapsulated IP tunnel—but the application is radically different. A consumer VPN ensures "privacy" for only the part of the traffic transferred between the consumer device and the VPN gateway; after that it is public again. But the consumer VPN is widespread and useful, so the industry now refers to both products as VPNs.

Limitations: Identity. The VPN does a decent job of encrypting traffic over public networks, but one weakness of the model traditionally is that IP addresses do not line up with authorization identities. There are two variants of this problem.

First, multiple users on a single computer all share an IP address. Network stacks are traditionally considered a computerwide service provided

by the operating-system kernel. Thus, any user on a computer can act as a VPN tunnel's IP address.


Second, if a VPN gateway is used to connect two subnets of machines together, there is no way to map the credentials used to establish the VPN to identify the machines on the subnet it routes. You have expanded your network to include all of the network on the other end of the VPN. If you don't manage both networks yourself, then you have just created a new network administrator.

These limitations are frustrating because it means a traditional corporate VPN does not provide user security. All the effort put into managing credentials on both ends of the VPN has to be done again one layer up, between users over the VPN.


Disillusionment: BeyondCorp and zero trust. The scale of the Internet strains the security model of the second-age VPN. Smaller organizations using the VPN as a cheaper leased line, then letting their traveling workers use that leased line from any hotel room, face a growing problem: With far fewer resources than a large organization, the small organization needs to secure an access point to its network that is getting ever cheaper to attack. As the number of Internet users went from millions to billions, corporate VPN gateways went from thousands to (potentially) millions. Each of these gateways runs one of a handful of implementations, uses common usernames, lacks two-factor authentication, and has common unmanaged passwords. An attack written to find an exploit on some fraction of existing targets becomes more cost effective as the number of targets grows.

The growing threat interacts poorly with the “eggshell” security model (hard exterior, soft gooey interior) of traditional corporate VPNs: the idea that the VPN keeps your network safe, so you can be lax about what you transmit. As VPNs become more cost-effective targets as their numbers grow, corporate VPNs get compromised more often.

This has led several security experts to call for VPNs to be dismantled.¹⁰ New security models have been proposed to replace the VPN. Two notable approaches are BeyondCorp,¹² a proj-



IPSec does everything, and everything it does is configurable. It has two modes of operation—tunnel and transport—multiple cryptographic suites, and multiple independent implementations.



ect by Google to secure its corporate network infrastructure; and a developing industry idea known as the *zero trust network*.

BeyondCorp and zero trust have a lot of conceptual overlap and can best be summarized as applications of the venerable end-to-end principle from computer networking. Specifically, when any two services communicate, each service must mutually authenticate who it is talking to and ensure that service is authorized to communicate.

This concept of mutual authentication is incompatible with the traditional corporate VPN “gateway” or “concentrator,” a device that sits on a network and routes all traffic through it, because the devices on either side of the gateway cannot be sure they are who they say they are.

The second age of the corporate VPN is coming to an end. Its security model is incompatible with modern Internet scale. It is too soon to declare the death of the VPN, however. Consumer VPNs continue to be useful tools, and while the traditional way of configuring corporate VPNs is clearly over, the underlying concept of encapsulating an IP packet still works. Put another way, the corporate application needs rethinking, but there still may be a use for the technology.

The Third Age: Single-Use Devices and Virtual Network Namespaces

The big exciting VPN development of the past few years is WireGuard,⁴ a completely new implementation of IP encapsulation using the latest in cryptographic algorithms and principles.

WireGuard, the creation of Jason A. Donenfeld, is built on top of the cryptographic primitives *curve25519* and *chacha20*.¹³ The protocol creates a tunnel between two equal peers, each identified with public/private key pairs rather than the common client-server architecture of VPNs with gateways and concentrators. It adopts handshake techniques and principles of the Noise Protocol⁹ to make it practically impossible for adversaries even to know a machine is running a WireGuard endpoint. There is no standard port to scan for on a network.

What makes WireGuard radical, however, is not its adoption of the very latest in cryptographic algorithms

(which many would consider a classic virtue of developing a product from scratch). WireGuard is radically simple. It has only one cryptographic suite. There is no version-negotiation phase of the protocol, where multiple implementations try to agree on how to talk. It tries to be exactly one thing: a secure encrypted tunnel between two endpoints.


In a world where networking software tries to be everything to everyone, where configuration languages need their own independent standards committee,² WireGuard is a breath of fresh air. It simplifies encrypted tunnels to the point where you can stop thinking of it as the final product, and instead consider it a fundamental abstraction around which software and networks can be designed.

Single-use devices and zero-trust VPNs. In addition to an exciting new VPN protocol, there have been fundamental shifts in the way computers are used that create new uses for VPNs and solve old security issues.


The most significant technical shift in network computing of the past decade has been the rise of the single-use device. This is driven by two changes to computing.

First, interactive user devices are effectively single-user operating systems because they allow only a single user at a time. The most extreme example of this is iOS—more than a billion active devices—which does not even support multiple user accounts on a single device. But even more traditional desktops and laptops have typically only one logged-in user. Rarer devices that support fast user switching can be configured so the VPN disconnects as one user and connects as another on a switch. The large shared Unix mini-computers with terminals are now interesting hobby projects, not typical corporate setups.

The second change is the near-universal virtualization of servers with virtual machines or container technology such as Linux namespaces. There are several forces driving this change and several ways it is achieved, but the result is the same: A server ends up running on a multiuser, multitasking Unix operating system with effectively a single-purpose process and one user.



The growing number of end-user devices and a new layer of virtualization in datacenters has subtly but profoundly changed how the VPN abstraction fits into networking.



In both of these cases, the result is that the operating system's virtual tunnel IP addresses now line up with service identities used for authorization. On end-user devices an IP address is a user, and in datacenters each service instance has its own IP. With WireGuard ensuring every packet with a particular IP source is cryptographically linked to a verifiable identity, we can start safely making statements such as, "Address *a* is user *u*," which simplifies software development. Tailscale⁸ is an implementation of a VPN-identified network built on WireGuard.

The growing number of end-user devices and a new layer of virtualization in datacenters has subtly but profoundly changed how the VPN abstraction fits into networking. With a little care in a modern environment, the traditionally awkward and unhelpful security model of the VPN suddenly fits perfectly and solves problems instead of creating them. This is what makes the third age of the VPN so exciting: The clumsy 1990s child, a millennial often dismissed as awkward and out of place, is suddenly making computing easier and better. **C**

References

1. Aboba, B. et al. Review of roaming implementation. IETF Network Working Group, 1997; <https://tools.ietf.org/html/rfc2194>.
2. Ben-Kiki, O., Evans, C., dot Net, I. YAML specification index, 2009; <https://yaml.org/spec/>.
3. Digital Equipment Corporation: nineteen fifty-seven to the present, 1978, 53; <https://www.computerhistory.org/pdp-1/8a9cb4c9f949fbb3e577016d174499ca/>.
4. Finley, K. WireGuard gives Linux a faster, more secure VPN. *Wired* (Mar. 2, 2020); <https://www.wired.com/story/wireguard-gives-linux-faster-secure-vpn/>.
5. Hamzeh, K. et al. Point-to-Point Tunneling Protocol. IETF Network Working Group, 1999; <https://tools.ietf.org/html/rfc2637>.
6. Ioannidis, J., Blaze, M. The swIPe security protocol. Internet draft, 1993; <https://www.mattblaze.org/papers/swipe.id.txt>.
7. Kent, S., Atkinson, R. Security architecture for the Internet Protocol. IETF Network Working Group, 1998; <https://tools.ietf.org/html/rfc2401>.
8. Pennarun, A. How Tailscale works. Tailscale, 2020; <https://tailscale.com/blog/how-tailscale-works/>.
9. Perrin, T. The Noise Protocol Framework, 2018; <http://noiseprotocol.org/noise.pdf>.
10. Sullivan, P. The death of the VPN—it's time to say goodbye. *SC Magazine* (Mar. 21, 2019); <https://www.scmagazine.com/home/opinion/the-death-of-the-vpn-its-time-to-say-goodbye/>.
11. Valencia, A. et al. Cisco Layer 2 Forwarding Protocol. IETF Network Working Group, 1998; <https://tools.ietf.org/html/rfc2341>.
12. Ward, R., Beyer, B. BeyondCorp: A new approach to enterprise security. *login*; 39, 6 (2014), 6–11; <https://research.google/pubs/pub43231/>.
13. WireGuard; <https://www.wireguard.com/protocol/>.

David Crawshaw is cofounder and CTO of Tailscale. Previously, he worked on a variety of software projects, including the Go programming language.

Copyright held by owner/author.
Publication rights licensed to ACM.

volume
01

number
01

FIRST
ISSUE
PUBLISHED

ACM Transactions on Internet of Things
is now available in
the ACM Digital Library



ACM Transactions on Internet of Things (TIOT) publishes novel research contributions and experience reports in several research domains whose synergy and interrelations enable the IoT vision. TIOT focuses on system designs, end-to-end architectures, and enabling technologies, and on publishing results and insights corroborated by a strong experimental component.

What does it mean to be fair?

BY SORELLE A. FRIEDLER, CARLOS SCHEIDEGGER,
AND SURESH VENKATASUBRAMANIAN

The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making

AUTOMATED DECISION-MAKING SYSTEMS (often machine learning-based) now commonly determine criminal sentences, hiring choices, and loan applications. This widespread deployment is concerning, since these systems have the potential to discriminate against people based on their demographic characteristics. Current sentencing risk assessments are racially biased,⁴ and job advertisements discriminate on gender.⁸ These concerns have led to an explosive growth in fairness-aware machine learning, a field that aims to enable algorithmic systems that are fair by design.

To design fair systems, we must agree precisely on what it means to be fair. One such definition is

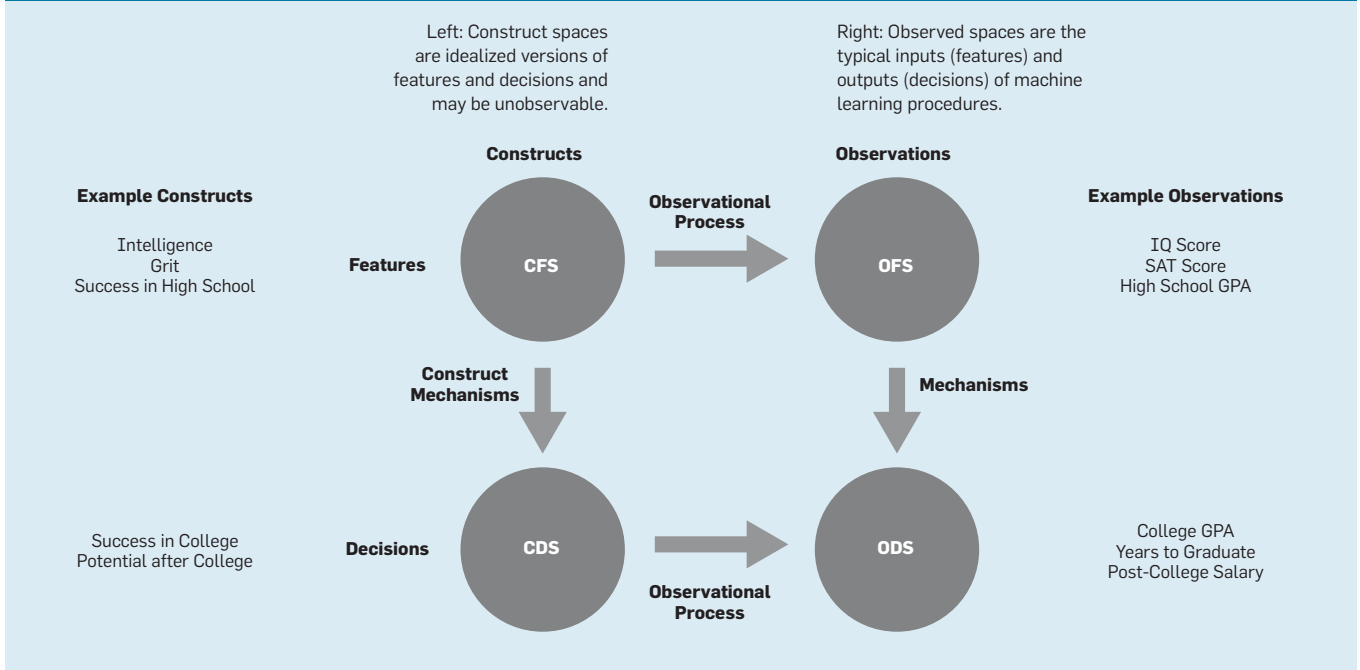
individual fairness:¹⁰ individuals who are similar (with respect to some task) should be treated similarly (with respect to that task). Simultaneously, a different definition states that demographic groups should, on the whole, receive similar decisions. This *group fairness* definition is inspired by civil rights law in the U.S.^{5,11} and U.K.²¹ Other definitions state that fair systems should err evenly across demographic groups.^{7,13,24} Many of these definitions have been incorporated into machine learning pipelines.^{1,6,11,16,25}

In this article, we introduce a framework for understanding these different definitions of fairness and how they relate to each other. Crucially, our framework shows these definitions and their implementations correspond to different axiomatic beliefs about the world. We present two such *worldviews* and will show they are fundamentally incompatible. First, one can believe the observation processes that generate data for machine learning are structurally biased. This belief provides a justification for seeking non-discrimination. When one believes that demographic groups are, on the whole, fundamentally similar, group fairness mechanisms successfully guarantee the top-level goal of non-discrimination: similar groups receiving similar treatment. Alternatively, one can assume the observed data generally re-

» key insights

- **The world is structurally biased and makes structurally biased data. Observation is a process. When we create data, we choose what to look for.**
- **Every automated system encodes a value judgment. Accepting training data as given implies structural bias does not appear in the data and that replicating the data as given would be just.**
- **Different value judgments can require satisfying contradicting fairness properties each leading to different societal outcomes.**
- **Researchers and practitioners must document data collection processes, worldviews, and value assumptions.**
- **Value decisions must come from domain experts and affected populations; data scientists should listen to them in order to build in values that lead to justice.**

Figure 1. Our model of algorithmic decision making involves transformations among four spaces.



flects the true underlying reality about differences between people. These worldviews are in conflict; a single algorithm cannot satisfy either definition of fairness under both worldviews. Thus, researchers and practitioners ought to be intentional and explicit about worldviews and value assumptions: the systems they design will always encode some belief about the world.

An Example

We illustrate the practice of fairness in decision making with the example of a college admissions process. We often think of this process as starting with the input provided by application materials and ending with an admittance decision. Here, we will take a broader view of the process, including the goals of the admissions office and an assessment of the resulting decisions. In this broader view, the first step of the process is determining a set of idealized features to be used in an admissions decision.

For example, some consider personal qualities such as self-control, growth mindset, and grit to be determining factors in later success.³ *Grit* is roughly defined as an individual’s ability to demonstrate passion, perseverance, and resilience toward their chosen goal. An admissions committee could decide to use grit as an idealized predictor. However, grit (and other idealized features) are not directly

observable; they are measured indirectly and imprecisely through self-reported surveys and other proxies.⁹

Similarly, when considering admissions decisions, it is also important to determine what an idealized decision should be predicting. An admissions office might decide that decisions should be made based on the predicted potential of an applicant. Since “potential” is unobservable, systems might use more directly measurable—and more problematic—features such as college GPA upon graduation.

This college admissions example demonstrates the basic pipeline in human decision-making systems, which is also mimicked in algorithmic systems. We decide on idealized features, and measure observed, possibly flawed versions of these features. We determine an idealized prediction goal, and measure observed features to predict an observed goal. We next formalize this decision-making framework.

Spaces: Construct vs. Observed and Features vs. Decisions

We model an algorithm making decisions about individuals as a mapping from a space of information about people, which we will call a *feature space*, to a space of decisions, which we will call a *decision space*. We assume each space is a collection of information about people endowed with a *dis-*

tance metric (specifically, a function of pairs of elements that satisfies reflexivity, symmetry, and the triangle inequality). This reflects that a process for discovering mappings from features to decisions usually exploits the *geometry* of these spaces.

We introduce two types of spaces: *construct spaces* and *observed spaces*. Construct spaces contain an idealized representation of information about people and decisions. These spaces may include unmeasurable “constructs” (for example, grit). Observed spaces contain the results of an observational process that maps information about people or decisions to measurable spaces of inputs or outputs (for example, the results of self-reported surveys designed to measure grit). An observational process can be noisy, including additional information not found in the associated construct space, missing information, or even containing a large distance skew in the mapping. Observational processes don’t have to maintain information that is useful for the decision-making task.

These two distinctions—between feature and decision spaces and between construct and observed spaces—naturally give rise to four spaces that we claim are necessary for analyzing the fairness of a decision-making procedure (as illustrated in Figure 1):

The Construct Feature Space (CFS) is the space representing the “desired”

or “true” collection of information about people to use as input to a decision-making procedure. For example, this includes features like intelligence or grit for college admission.

The Observed Feature Space (OFS) is the space containing the observed information about people, generated by an observational process $g : CFS \rightarrow OFS$ that generates an entity $\hat{p} = g(p)$ from a person $p \in CFS$. For example, this includes the results of standardized tests or personal essays.

The Construct Decision Space (CDS) is the space representing the idealized outcomes of a decision-making procedure. For example, this includes how well a student will do in college.

The Observed Decision Space (ODS) is the space containing the per-person observed decisions from a concrete decision-making procedure, generated by an observational process mapping $CDS \rightarrow ODS$. For example, this includes the GPA of a student after their freshman year.

To understand the interactions between these spaces, we start with a prediction task, determine an idealized decision goal, posit features that seem to control the decision, and then imagine ways of measuring those features and decisions. Explicitly considering the existence of the construct space is rare in practice; we argue that explicit goals and assumptions are necessary when considering fairness. It is worth emphasizing here that the construct spaces represent our best *current* understanding of the underlying factors involved in a task rather than some kind of Platonic universal ideal. They are therefore *contingent* on the current specific ideas and best practices about how to make the decision in the given context.

TL;DR

Constructs are the idealized features and decisions we wish we could use for decision-making.

Observed features and decisions are the measurable features and outcomes that are actually used to make decisions.

These **may be different**, and it is important to be explicit about the distinction.

Fairness and Non-Discrimination

Traditional data science and machine learning can be understood as

focusing on creating transformations between the observed feature and observed decision spaces. These mechanisms are used in real-world decision-making practices by taking observed data as input. On the other hand, fairness is defined as a property of an idealized *construct mechanism* that maps individuals to construct decisions based on their construct features. The goal of algorithmic fairness is to develop real-world mechanisms that match the decisions generated by these construct mechanisms. In order to discuss these fairness-aware mechanisms further, we first describe different notions of fairness within our framework.

Individual fairness. Since fairness is an idealized property operating based on underlying and potentially unobservable information about people, it is most natural to define it within the construct spaces. The definition of fairness is task specific and prescribes desirable (potentially unobservable) outcomes for a task in the construct decision space. Since the solution to a task is a mapping from the construct feature space to the construct decision space, a definition of fairness should describe the properties of such a construct mechanism. Inspired by the fairness definition due to Dwork et al.,²⁰ we define individual fairness as follows:

Individual fairness. Individuals who are similar (with respect to the task) in the *CFS* should receive similar decisions in the *CDS*.

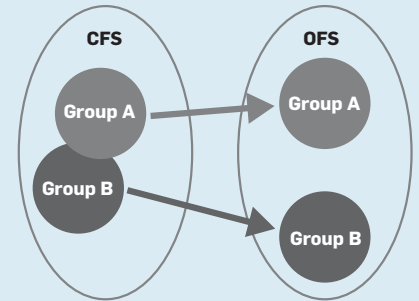
TL;DR

Individual fairness is the goal of giving similar individuals similar decisions.

Non-discrimination. While our fairness definition focuses on the individual, there are often groups that share characteristics (such as race, gender, and so on) and fairness should be considered with respect to these group characteristics (or combinations of them) as well. Group membership is often defined on innate, culturally defined, immutable characteristics, or those protected for historical reasons. It is often considered unacceptable (and sometimes illegal) to use group membership as part

Figure 2. An illustration of group skew in the mapping between the feature spaces.

The mapping moves the groups further apart in the observed space than they were in the construct space, increasing the inter-group distance while maintaining the intra-group distance.



of a decision-making process. Thus, we define non-discrimination as follows:

Non-discrimination. Groups who are similar (with respect to the task) in the *CFS* should, as a whole, receive similar decisions in the *CDS*.

In this work, consider group membership as a characteristic of an individual; thus, each of the four spaces admits a partition into groups, induced by the group memberships of individuals represented in these spaces.

Discrimination manifests itself in unequal treatment of groups. To quantify this, we first introduce the notion of *group skew*. Given two spaces and their associated group partitioning, the group skew of the mapping between the spaces is the extent to which the groups are, as a whole, mapped differently from each other. An illustration of group skew between the feature spaces is given in Figure 2. The goal in any formalization of group skew is to capture the relative difference in the mappings of groups with respect to each other, rather than (for example) a scaling transformation that transforms all groups the same way. This can be thought of as quantifying any difference in treatment based on group membership.^a Thus, nondiscrimination is defined as the lack of group skew in the mapping from *CFS* to *CDS*. This notion attempts to capture the idea of fair treatment of groups.

^a We introduce one possible geometric formalization of group skew in Friedler et al.¹²

TL;DR

Non-discrimination is the goal of giving similar groups on the whole similar decisions.

While non-discrimination shares many characteristics with the notions of group fairness that previous work has studied, an important distinction that we introduce here is that non-discrimination is defined as a property of the mapping between the *construct* spaces, while group fairness is generally defined in practice via *group fairness mechanisms* that restrict mappings between the *observed* feature and decision spaces. The ideas we develop next will allow us to understand that mechanisms that guarantee group fairness notions are generally doing so with the goal of guaranteeing non-discrimination, but formalizing the relationship between these two notions requires axiomatic assumptions about the world.

Worldviews and Assumptions

Fairness goals are defined as properties of construct mechanisms. Real-world decision making, however, must use mechanisms that map between the observed spaces. Thus, fair algorithm designers are forced to make assumptions about the observational processes mapping from construct to observed spaces in order to make real-world decisions.

We now describe two such axiomatic assumptions that are motivated by associated worldviews. While we introduce these two axioms as different worldviews or belief systems, they can also be strategic choices. Roemer identifies the goal of such choices as ensuring that negative attributes due to an individual’s circumstances of birth or to random chance should not be held against them, while individuals should be held accountable for their effort and choices.²⁰ In our framework, this translates to a decision of which axiom to choose. In our college admissions example, this may be the difference between asserting the admissions process should serve as a social equalizer, so that, for example, applicants from different class backgrounds are admitted at approximately the same rate, or believing that features such as GPA and SAT scores accurately reflect effort and understanding.

Worldview: What you see is what you get. One worldview handles the uncertainty of the observational process mapping between any construct and observed space by asserting that the spaces are *essentially the same* with respect to the task.

What You See Is What You Get (WYSIWYG): The observational process between a given construct space and observed space maintains the relative position of individuals with respect to the task.

It is common in data science to directly use any observed data that is available; doing so without modification or further evaluation is an adoption of the WYSIWYG worldview and assumption. Importantly, a consequence of understanding that WYSIWYG is an axiomatic assumption is that we can reevaluate this assumption.

Worldview: Structural bias. What if the construct space is not accurately represented by the observed space? In many real-world applications, the transformation from construct to observed space is non-uniform in a societally biased way. To capture this idea, we return to the notion of group skew from the previous section. We define structural bias as group skew between construct and observed spaces, that is, an observational process that treats groups differently. Such a process is illustrated in Figure 2.

In the cases where observed data is believed to suffer from structural bias, a common fairness goal is non-discrimination—a goal that aims to make sure that the group skew that is present in the observed data is not found in the resulting decisions. Unfortunately, non-discrimination is difficult to achieve directly since it is defined in terms of the construct spaces, and the existence of structural bias precludes us from using the observed spaces as a reasonable representation of the construct spaces (unlike the WYSIWYG worldview).

Instead, a common underlying assumption of this worldview is that in one or both of the construct spaces *all groups look essentially the same*. It asserts there are no innate differences between demographic groups. There may still be variation between individuals within the group, but the assumption is that on the whole, for example, as a distribution, the groups are essentially the same.

We’re All Equal (WAE). Within a given construct space all groups are essentially the same.

This axiom could be applied to either or both of the construct spaces (*CFS* and *CDS*). It appears implicitly in much of the literature on statistical discrimination and disparate impact.

There is an alternate interpretation of this axiom: the groups are not necessarily equal, but for the purposes of the decision-making process, we would prefer to treat them as if they were. In this interpretation,²⁰ any difference in the groups’ performance (for example, academic achievement) is due to factors outside their individual control (for example, the quality of their neighborhood school) and should not be taken into account in the decision-making process. This interpretation has the same mathematical outcome as if groups are assumed equal, and thus a single axiom covers both these interpretations.

We reiterate that structural bias—the way in which observations are systematically distorted—is separate from the WAE axiom, which is a device that allows us to interpret observed skew as a measure of structural bias.

TL;DR

Any attempt to design fair decision making is **forced to make assumptions** about the observational process and/or construct space. There are two main such assumptions:

Work that **uses the observed data directly** is making a WYSIWYG assumptions; and,

Work that attempts to guarantee statistical parity and other **group fairness notions as a measure** is making a WAE assumption.

Later, we will explore the ways that different works have made these assumptions further.

Consequences

We now sketch some consequences of attempts to achieve individual fairness and non-discrimination under different worldviews. A more detailed analysis can be found in our extended work;¹² other authors have also started to build on the framework we lay out.²³

Mechanisms achieve the goals of their worldviews. How can individual fairness be achieved? *Individual fairness mechanisms* are algorithms that guarantee that individuals who are similar in the observed feature space receive similar decisions in the observed decision space. Under WYSIWYG assumptions on both the feature and decision observational processes, individual fairness mechanisms can be shown to guarantee individual fairness (via function composition).

Group fairness mechanisms ensure that groups are mapped to, on the whole, similar decisions in the observed decision space. Under a WAE assumption (applied to the *CFS*), group fairness mechanisms can be shown to guarantee non-discrimination since groups are assumed to be essentially the same in the construct feature space and the mechanism guarantees that this is enforced in the mapping to the observed decision space.

TL;DR

Under a **WYSIWYG** assumption, **individual fairness** can be guaranteed.

Under a **WAE** assumption, **non-discrimination** can be guaranteed.

Conflicting worldviews necessitate different mechanisms. Do mechanisms exist that can guarantee individual fairness or non-discrimination under both worldviews?

Unfortunately, WYSIWYG appears to be crucial to ensuring individual fairness: if there is structural bias in the decision pipeline, no mechanism can guarantee individual fairness. Fairness can only be achieved under the WYSIWYG worldview using an individual fairness mechanism and using a group fairness mechanism will be *unfair* within this worldview.

What about non-discrimination? Unfortunately, another counterexample shows these mechanisms are not agnostic to worldview. Suppose that the construct and observed decision spaces are the same and that two groups are very far apart in the *CFS* with images in the *OFS* that are even further apart. Applying an individual fairness mechanism to the *OFS* will result in decisions that preference the group that performed better

Researchers and practitioners ought to be intentional and explicit about worldviews and value assumptions—the systems they design will always encode some belief about the world.

with respect to the task in the *CFS* more than is warranted compared to the other group; this is discriminatory.

Choice in mechanism must thus be tied to an explicit choice in worldview. Under a WYSIWYG worldview, only individual fairness mechanisms achieve fairness (and group fairness mechanisms are unfair). Under a structural bias worldview, only group fairness mechanisms achieve non-discrimination (and individual fairness mechanisms are discriminatory).

TL;DR

Fairness-aware algorithms cannot guarantee fairness or non-discrimination under *both* the WYSIWYG and structural bias worldviews. Choice in algorithms must be tied to an **explicit choice in worldview**.

Placing Literature in Context

Our framework lets us analyze existing literature in fairness (for broader surveys, see Romei et al.²¹ and Zliobaite²⁷) to see what axiomatic positions different solutions might implicitly be taking. For this analysis, we distinguish papers that propose new fairness measures and/or interventions from the smaller number of papers that provide a metanalysis of fairness definitions.

Fairness measures and algorithms.

Our findings are twofold. First, we find we can categorize existing work based on fairness *measure* and associated assumption on the *decision spaces*. Measures that assume that existing decisions are correct and optimize fairness conditioned on that assumption adopt the WYSIWYG axiom between decision spaces. Measures that are open to changing the observed decisions in the data adopt the WAE axiom. Second, once we categorize measures based on the decision space axiomatic choice, we can categorize algorithms based on axioms governing the feature spaces. Algorithms that work to change the data representation adopt a viewpoint that the data may not be correct, and generally do this according to the WAE axiom. Algorithms that make no change to the data before optimizing for a measure implicitly make the WYSIWYG axiomatic assumption between feature spaces.

Note that it is not a contradiction to have an algorithm that, based on its measure, assumes WAE in the construct decision space and WYSIWYG

between the feature spaces. In fact, many algorithms make these dual assumptions in practice.

Early work on non-discrimination that initiated the study of fairness-aware data mining considered the difference in outcomes between groups. Specifically, let $Pr[C = \text{Yes} | G = 0]$ be the probability of people in the unprivileged group receiving a positive classification and $Pr[C = \text{YES} | G = 1]$ be the probability of people in the privileged group receiving a positive classification. Calders and Verwer⁶ introduce the idea of a *discrimination score* defined as $Pr[C = \text{YES} | G = 1] - Pr[C = \text{YES} | G = 0]$. Their goal, and the goal of much subsequent work also focusing on this measure,^{14,16,22,26} was to bring this difference to zero. The assumption here is that groups should, as a whole, receive similar outcomes. The implicit assumption is that the original decisions received by the groups (that is, the decisions used for training) may not be correct if this difference is not small. This reflects an underlying WAE axiom in the construct decision space. The four-fifths rule for disparate impact^{5,11,25} focuses on a similar measure (taking the ratio instead of the difference) and also assumes the WAE axiom.

A 2016 ProPublica study⁴ examined the predicted risk scores assigned to defendants by the COMPAS algorithm and found that Black defendants were about twice as likely to receive incorrect high-risk scores (bad errors), while White defendants were about twice as likely to receive incorrect low risk scores (good errors). This inspired the development of measures for equalizing the group conditioned error rates of algorithms (termed “equal odds”¹³ or “disparate mistreatment”²⁴), with the idea that different groups should receive the same impact of the algorithm conditioned on their outcomes. This implicitly assumes the observed outcomes (observed decision space) reflect true decisions, that is, these measures assume the WYSIWYG axiom between the decision spaces. It is interesting that while these classes of measures are all considered group fairness measures, they make different assumptions about the decision spaces.

Axiomatic assumptions about feature spaces are determined by the choice of algorithm. Some works attempt to ensure non-discrimination by modifying the decision algorithm^{6,16} while others



Discrimination manifests itself in unequal treatment of groups.



change the outcomes after the decision has been drafted.¹⁵ Even though these algorithms try to ensure non-discrimination and assume the WAE axiom in the construct decision space, they implicitly assume the WYSIWYG axiom between the feature spaces by using training data without modification. Algorithms for achieving fairness on group-conditioned error rates^{13,24} focus on constraining this measure using the observed data as given, so these algorithms assume WYSIWYG between the feature spaces. Given that these algorithms focused on the group-conditioned error rate also assume the WYSIWYG between decision spaces, we claim that they adopt the WYSIWYG worldview and not a structural bias worldview.

Other algorithms perform preprocessing on the training data.^{11,14,19,26} These works can be seen as attempting to reconstruct the construct feature space and make decisions based on that hypothesized reality under the WAE assumption.

We turn now to Dwork et al.’s individual fairness definition:¹⁰ two individuals who are similar should receive similar outcomes. Dwork et al. emphasize that determining whether two individuals are similar with respect to the task is critical and assume such a metric is given. In light of the formalization of the construct spaces, we note that the metric discussed by Dwork et al. is the distance in a combined construct space including both features and decisions. As described by Dwork et al., the metric is not known. We claim that in practice this lack of knowledge is resolved by the axiomatic assumption of either WYSIWYG or WAE, and since the focus is on individual fairness and not on groups, the WYSIWYG assumption is usually made between both feature and decision spaces.

TL;DR

Fairness **measure** choices encode assumptions about the **decision** spaces. Parity-focused notions (for example, disparate impact) assume WAE. Error rate balance assumes WYSIWYG.

Intervention **algorithm** choices encode assumptions about the **feature** spaces. Representational approaches assume WAE. In-processing and post-processing approaches assume WYSIWYG.

Fairness meta-analyses. Further examination of group fairness measures prompted the discovery of the mutual incompatibility of error rate balance (for both positive and negative classifications) and equality of per-group calibration (a measure indicating if a score is correctly predicting graded outcomes). These constraints cannot be simultaneously achieved unless the classifier is perfect or the base rates per group are equal.^{7,17} Since these measures naturally make a WYSIWYG assumption between the decision spaces, this impossibility result only holds under this axiom. In fact, the case under which it no longer holds—the base rates per group being equal—is one possible codification of the WAE axiom in the construct decision space.

TL;DR

Fairness impossibility results^{7,17} hold under the WYSIWYG axiom between the decision spaces. These results do not hold under the WAE axiom between decision spaces.

Meta-analyses should make their axiomatic assumptions explicit and consider both measures and algorithms.


Discussion and Conclusion

Our main claim in this work is that discussions about fairness algorithms and measures should make explicit the implicit assumptions about the world being modeled. The focus by traditional data science techniques on the observed feature and decision spaces obscures these important axiomatic issues. The default assumption in these traditional data science and machine learning domains is the WYSIWYG assumption; the data is taken as given and fully representative of the implicit construct spaces. In this work, we highlight that this WYSIWYG assumption should be made purposeful and explicitly.

When considering fairness-aware algorithms applied to a specific domain, all assumptions are not equally reasonable. There is extensive social science literature demonstrating the existence of structural bias in criminal justice,² education,¹⁸ and other fairness-critical domains. In these domains, it is not reasonable to make the

WYSIWYG assumption. Data science practitioners must work with domain experts and those impacted by resulting decisions to understand what assumptions are reasonable in a given context before developing and deploying fair mechanisms; without this work, incorrect assumptions could lead to unfair mechanisms.

Additionally, our framework suggests ways in which the current discussion of fairness measures is misleading. First, group and individual notions of fairness reflect fundamentally different underlying goals and are not mechanisms toward the same outcome. Second, group notions of fairness differ based on their implicit axiomatic assumptions: mathematical incompatibilities should be viewed as a formal statement of this more philosophical difference. And finally, and perhaps most importantly, comparing definitions of fairness is incomplete without also discussing the deployed interventions: it is the combination of measure and algorithm that describes a fully specified worldview in which the system operates.

Acknowledgments. This research was funded in part by the NSF under grants IIS-1251049, CNS-1302688, IIS-1513651, IIS-1633724, and IIS-1633387. Thanks to the attendees at the Dagstuhl Workshop on Data Responsibly for their helpful comments, and to Cong Yu, Michael Hay, Nicholas Diakopoulos and Solon Barocas. We also thank Tionney Nix, Tosin Alliyu, Andrew Selbst, danah boyd, Karen Levy, Seda Gürses, Michael Ekstrand, Vivek Srikumar, and Hannah Sassaman and the community at Data & Society. 

References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th Intern. Conf. on Machine Learning 80*. J. Dy and A. Krause, (Eds.). PMLR, (Stockholmsmässan, Stockholm Sweden, 2018), 60–69; <http://proceedings.mlr.press/v80/agarwal18a.html>
2. Alexander, M. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2012.
3. Almlund, M., Duckworth, A., Heckman, J., and Kautz, T. *Personality psychology and economics*. Technical Report w16822. NBER Working Paper Series. National Bureau of Economic Research, Cambridge, MA, 2011.
4. Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica* (May 23, 2016).
5. Barocas, S. and Selbst, A. Big data's disparate impact. *California Law Review* 104, 671. (2016).
6. Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Min Knowl Disc* 21 (2010), 277–292.
7. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

8. Datta, A., Tschantz, M., and Datta, A. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proceedings on Privacy Enhancing Technologies I* (2015), 92 – 112.
9. Duckworth, A., Peterson, C., Matthews, M., and Kelly, D. Grit: Perseverance and passion for long-term goals. *J. Personality and Social Psychology* 92, 6 (2007), 1087–1101.
10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf.* (2012), 214–226.
11. Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, (2015), 259–268.
12. Friedler, S., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness; *arXiv:1609.07236* (2016).
13. Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016, 3315–3323.
14. Kamiran, F. and Calders, T. Classifying without discriminating. In *Proceedings of the 2nd Intern. Conf. Computer, Control and Communication*. IEEE, (2009), 1–6.
15. Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. *ICDM*, (2012), 924–929.
16. Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases* (2012), 35–50.
17. Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science*, (2017), 43:1–43:23.
18. Kozol, J. *The Shame of the Nation: The Restoration of Apartheid Schooling in America*. Broadway Books, 2006.
19. Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *Proceedings of the 35th Intern. Conf. on Machine Learning 80*. J. Dy and A. Krause (Eds.), PMLR, (2018), 3384–3393; <http://proceedings.mlr.press/v80/madras18a.html>
20. Roemer, J. *Equality of Opportunity*. Harvard University Press, 1998.
21. Romei, A. and Ruggieri, S. A Multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review* (Apr. 3, 2013), 1–57.
22. Ruggieri, S. Using t-closeness anonymity to control for nondiscrimination. *Transactions on Data Privacy* 7 (2014), 99–129.
23. Yeom, S. and Tschantz, M. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. *CoRR* abs/1808.08619 (2018). [arXiv:1808.08619](http://arxiv.org/abs/1808.08619)
24. Zafar, M., Valera, I., Rodriguez, M., and Gummadi, K. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of WWW*, (2017), 1171–1180.
25. Zafar, M., Valera, I., Roriguez, M., and Gummadi, K. 2017. Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics*, (2017), 962–970.
26. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *Proceedings of ICML*, (2013), 325–333.
27. Zliobaite, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.

Sorelle A. Friedler (sorelle@cs.haverford.edu) is an associate professor of computer science at Haverford College, Haverford, PA, USA.

Carlos Scheidegger (cscheid@email.arizona.edu) is an associate professor of computer science at the University of Arizona, Tucson, AZ, USA.

Suresh Venkatasubramanian (suresh@cs.utah.edu) is a professor of computer science at the University of Utah, Salt Lake City, UT, USA.

Copyright held by authors/owners. Publication rights licensed to ACM.

The promise and the challenges of the first industry-supported language to master the trade-off between safety and control.

BY RALF JUNG, JACQUES-HENRI JOURDAN, ROBERT KREBBERS, AND DEREK DREYER

Safe Systems Programming in Rust

THERE IS A longstanding tension in programming language design between two seemingly irreconcilable desiderata.

► **Safety.** We want strong type systems that rule out large classes of bugs statically. We want automatic memory management. We want data encapsulation, so that we can enforce invariants on the private representations of objects and be sure that they will not be broken by untrusted code.

► **Control.** At least for “systems programming” applications like Web browsers, operating systems, or game engines, where performance or resource constraints are a primary concern, we want to determine the byte-level representation of data. We want to optimize the time and space usage of our programs using low-level programming techniques. We want access to the “bare metal” when we need it.

Sadly, as conventional wisdom goes, we cannot have everything we want. Languages such as Java give us strong

safety, but it comes at the expense of control. As a result, for many systems programming applications, the only realistic option is to use a language like C or C++ that provides fine-grained control over resource management. However, this control comes at a steep cost. For example, Microsoft recently reported that 70% of the security vulnerabilities they fix are due to memory safety violations,³³ precisely the type of bugs that strong type systems were designed to rule out. Likewise, Mozilla reports that the vast majority of critical bugs they find in Firefox are memory related.¹⁶ If only there were a way to somehow get the best of both worlds: a *safe systems programming language with control...*

Enter **Rust**. Sponsored by Mozilla and developed actively over the past decade by a large and diverse community of contributors, Rust supports many common low-level programming idioms and APIs derived from modern C++. However, unlike C++, Rust enforces the safe usage of these APIs with a strong static type system.

» key insights

- **Rust is the first industry-supported programming language to overcome the longstanding trade-off between the safety guarantees of higher-level languages and the control over resource management provided by lower-level “systems programming” languages.**
- **It tackles this challenge using a strong type system based on the ideas of ownership and borrowing, which statically prohibits the mutation of shared state. This approach enables many common systems programming pitfalls to be detected at compile time.**
- **There are a number of data types whose implementations fundamentally depend on shared mutable state and thus cannot be type-checked according to Rust’s strict ownership discipline. To support such data types, Rust embraces the judicious use of unsafe code encapsulated within safe APIs.**
- **The proof technique of semantic type soundness, together with advances in separation logic and machine-checked proof, has enabled us to begin building rigorous formal foundations for Rust as part of the RustBelt project.**



In particular, like Java, Rust protects programmers from memory safety violations (for example, “use-after-free” bugs). But Rust goes further by defending programmers against other, more insidious anomalies that no other mainstream language can prevent. For example, consider *data races*: unsynchronized accesses to shared memory (at least one of which is a write). Even though data races effectively constitute undefined (or weakly defined) behavior for concurrent code, most “safe” languages (such as Java and Go) permit them, and they are a reliable source of concurrency bugs.³⁵ In contrast, Rust’s type system rules out data races at compile time.

Rust has been steadily gaining in popularity, to the point that it is now being used internally by many major industrial software vendors (such as

Dropbox, Facebook, Amazon, and Cloudflare) and has topped Stack Overflow’s list of “most loved” programming languages for the past five years. Microsoft’s Security Response Center Team recently announced that it is actively exploring an investment in the use of Rust at Microsoft to stem the tide of security vulnerabilities in system software.^{8,25}

The design of Rust draws deeply from the wellspring of academic research on safe systems programming. In particular, the most distinctive feature of Rust’s design—in relation to other mainstream languages—is its adoption of an *ownership type system* (which in the academic literature is often referred to as an *affine* or *substructural* type system³⁶). Ownership type systems help the programmer enforce safe patterns of lower-level program-

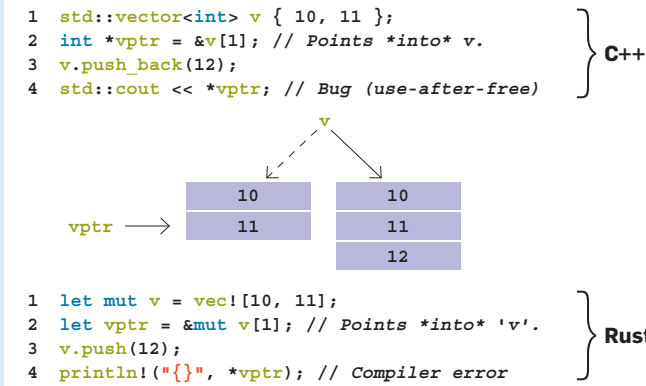
ming by placing restrictions on which *aliases* (references) to an object may be used to mutate it at any given point in the program’s execution.

However, Rust goes beyond the ownership type systems of prior work in at least two novel and exciting ways:

1. Rust employs the mechanisms of *borrowing* and *lifetimes*, which make it much easier to express common C++-style idioms and ensure they are used safely.

2. Rust also provides a rich set of APIs—for example, for concurrency abstractions, efficient data structures, and memory management—which fundamentally extend the power of the language by supporting more flexible combinations of aliasing and mutation than Rust’s core type system allows. Correspondingly, these APIs cannot be implemented within the safe fragment of Rust: rather, they internally make

Figure 1. Use-after-free bug in C++ and how the bug is prevented in Rust.



use of potentially *unsafe* C-style features of the language, but in a safely encapsulated way that is claimed not to disturb Rust’s language-level safety guarantees.

These aspects of Rust’s design are not only essential to its success—they also pose fundamental research questions about its semantics and safety guarantees that the programming languages community is just beginning to explore.

In this article, we begin by giving the reader a bird’s-eye view of the Rust programming language, with an emphasis on some of the essential features of Rust that set it apart from its contemporaries. Second, we describe the initial progress made in the RustBelt project, an ongoing project funded by the European Research Council (ERC), whose goal is to provide the first formal (and machine-checked) foundations for the safety claims of Rust. In so doing, we hope to inspire other members of the computer science research community to start paying closer attention to Rust and to help contribute to the development of this groundbreaking language.

Motivation: Pointer Invalidation in C++

To demonstrate the kind of memory safety problems that arise commonly in systems programming languages, let us consider the C++ code depicted at the top of Figure 1.

In the first line, this program creates a `std::vector` (a growable array) of integers. The initial contents of `v`, the two elements 10 and 11, are stored in a buffer in memory. In the second line, we create a pointer `vptr` that points into this buffer; specifically, it points to the

place where the second element (with current value 11) is stored. Now both `v` and `vptr` point to (overlapping parts of) the same buffer; we say that the two pointers are *aliasing*. In the third line, we push a new element to the end of `v`. The element 12 is added after 11 in the buffer backing `v`. If there is no more space for an additional element, a new buffer is allocated and all the existing elements are moved over. Let us assume this is what happens here. Why is this case interesting? Because `vptr` still points to the old buffer. In other words, adding a new element to `v` has turned `vptr` into a dangling pointer. This is possible because both pointers were aliasing: an action through a pointer (`v`) will in general also affect all its aliases (`vptr`). Figure 1 visualizes the entire situation.

The fact that `vptr` is now a dangling pointer becomes a problem in the fourth line. Here we load from `vptr`, and since it is a dangling pointer, this is a use-after-free bug.

In fact, the problem is common enough that one instance of it has its own name: *iterator invalidation*, which refers to the situation where an iterator (usually internally implemented with a pointer) gets invalidated because the data structure it iterates over is mutated during the iteration. It most commonly arises when one iterates over some container data structure in a loop, and indirectly, but accidentally, calls an operation that mutates the data structure. Notice that in practice the call to the operation that mutates the data structure (`push_back` in line 3 of our example) might be deeply nested

behind several layers of abstraction. In particular when code gets refactored or new features get added, it is often near impossible to determine if pushing to a certain vector will invalidate pointers elsewhere in the program that are going to be used again later.

Comparison with garbage-collected languages. Languages like Java, Go, and OCaml avoid use-after-free bugs using garbage collection: memory is only deallocated when it can no longer be used by the program. Thus, there can be no dangling pointers and no use-after-free.

One problem with garbage collection is that, to make it efficient, such languages generally do not permit *interior* pointers (that is, pointers *into* data structures). For example, arrays `int[]` in Java are represented similarly to `std::vector<int>` in C++ (except arrays in Java cannot be grown). However, unlike in C++, one can only *get* and *set* elements of a Java array, not take *references* to them. To make the elements themselves addressable, they need to be separate objects, references to which can then be stored in the array—that is, the elements need to be “boxed.” This sacrifices performance and control over memory layout in return for safety.

On top of that, garbage collection does not even properly solve the issue of iterator invalidation. Mutating a collection while iterating over it in Java cannot lead to dangling pointers, but it may lead to a `ConcurrentModificationException` being thrown at run time. Similarly, while Java *does* prevent security vulnerabilities caused by null pointer misuse, it does so with run-time checks that raise a `NullPointerException`. In both of these cases, while the result is clearly better than the corresponding undefined behavior of a C++ program, it still leaves a lot to be desired: instead of shipping incorrect code and then detecting issues at run time, we want to prevent the bugs from occurring in the first place.

Rust’s solution to pointer invalidation. In Rust, issues like iterator invalidation and null pointer misuse are detected statically, by the compiler—they lead to a compile-time error instead of a run-time exception. To explain how this works, consider the Rust translation of our C++ example at the bottom of Figure 1.

Like in the C++ version, there is a buffer in memory, and `vptr` points into the middle of that buffer (causing aliasing); `push` might reallocate the buffer, which leads to `vptr` becoming a dangling pointer, and that leads to a use-after-free in line 4.

But none of this happens; instead the compiler shows an error: “cannot borrow `v` as mutable more than once at a time.” We will come back to “borrowing” soon, but the key idea—the mechanism through which Rust achieves memory safety in the presence of pointers that point into a data structure—already becomes visible here: the type system enforces the discipline (with a notable exception that we will come to later) that *a reference is never both aliased and mutable at the same time*. This principle should sound familiar in the context of concurrency, and indeed Rust uses it to ensure the absence of data races as well. However, as our example that is rejected by the Rust compiler shows, the unrestricted combination of aliasing and mutation is a recipe for disaster even for sequential programs: in line 3, `vptr` and `v` alias (`v` is considered to point to all of its contents, which overlaps with `vptr`), and we are performing a mutation, which would lead to a memory access bug in line 4.

Ownership and Borrowing

The core mechanism through which Rust prevents uncontrolled aliasing is *ownership*. Memory in Rust always has a unique owner, as demonstrated by the example in Figure 2.

Here, we construct `v` similar to our first example, and then pass it to `consume`. Operationally, just like in C++, parameters are passed by value but the copy is shallow—pointers get copied but their pointee does not get duplicated. This means that `v` and `w` point to the same buffer in memory.

Such aliasing is a problem if `v` and `w` would *both* be used by the program, but an attempt to do so in line 6 leads to a compile-time error. This is because Rust considers ownership of `v` to have *moved* to `consume` as part of the call, meaning that `consume` can do whatever it desires with `w`, and the caller may no longer access the memory backing this vector at all.

Resource management. Ownership in Rust not only prevents memory

bugs—it also forms the core of Rust’s approach to memory management and, more generally, resource management. When a variable holding owned memory (for example, a variable of type `Vec<T>`, which owns the buffer in memory backing the vector) goes out of scope, we know for sure this memory will not be needed any more—so the compiler can automatically deallocate the memory at that point. To this end, the compiler transparently inserts *destructor* calls, just like in C++. For example, in the `consume` function, it is not actually necessary to call the destructor method (`drop`) explicitly. We could have just left the body of that function empty, and it would have automatically deallocated `w` itself.

As a consequence, Rust programmers rarely have to worry about memory management: it is largely automatic, despite the lack of a garbage collector. Moreover, the fact that memory management is also *static* (determined at compile time) yields enormous benefits: it helps not only to keep the maximal memory consumption down, but also to provide good *worst-case* latency in a reactive system such as a Web server. And on top of that, Rust’s approach generalizes beyond memory management: other resources like file descriptors, sockets, lock handles, and so on are handled with the same mechanism, so that Rust programmers do not have to worry, for instance, about closing files or releasing locks. Using destruc-

tors for automatic resource management was pioneered in the form of RAI (Resource Acquisition Is Initialization) in C++;³¹ the key difference in Rust is that the type system can statically ensure that resources do not get used any more after they have been destructed.

Mutable references. A strict ownership discipline is nice and simple, but unfortunately not very convenient to work with. Frequently, one wants to provide data to some function *temporarily*, but get it back when that function is done. For example, we want `v.push(12)` to grant `push` the privilege to mutate `v`, but we do not want it to *consume* the vector `v`.

In Rust, this is achieved through *borrowing*, which takes a lot of inspiration from prior work on *region types*.^{13,34} An example of borrowing is given in Figure 3. The function `add_something` takes an argument of type `&mut Vec<i32>`, which indicates a *mutable reference* to a `Vec<i32>`. Operationally, this acts just like a reference in C++, that is, the `Vec` is passed by reference. In the type system, this is interpreted as `add_something` *borrowing* ownership of the `Vec` from the caller.

The function `add_something` demonstrates what borrowing looks like in well-formed programs. To see why the compiler accepts that code while rejecting our pointer invalidation example from earlier, we have to introduce another concept: *lifetimes*. Just like in real life, when borrowing some-

Figure 2. Rust example: Moving ownership.

```
1 fn consume(w: Vec<i32>) {
2     drop(w); // deallocate vector
3 }
4 let v = vec![10, 11];
5 consume(v);
6 v.push(12); // Compiler error
```

Figure 3. Rust example: Mutable references.

```
1 fn add_something(v: &mut Vec<i32>) {
2     v.push(11);
3 }
4 let mut v = vec![10];
5 add_something(&mut v);
6 v.push(12); // Ok!
7 // v.push(12) is syntactic sugar for Vec::push(&mut v, 12)
```


thing, misunderstanding can be prevented by agreeing up front on how long something may be borrowed. So, when a reference gets created, it gets assigned a lifetime, which gets recorded in the full form of the reference type: `&'a mut T` for lifetime `'a`. The compiler ensures the reference (`v`, in our example) only gets used during that lifetime, and the referent does not get used again until the lifetime is over.

In our case, the lifetimes (which are all inferred by the compiler) just last for the duration of `add_something` and `Vec::push`, respectively. Never is `v` used while the lifetime of a previous borrow is still ongoing.

In contrast, Figure 4 shows the lifetimes inferred for the previous example from Figure 1. The lifetime `'a` of the borrow for `vp_ptr` starts in line 2 and goes on until line 4. It cannot be any shorter because `vp_ptr` gets used in line 4. However, this means that in line 3, `v` is used while an outstanding borrow exists, which is an error.

To summarize: whenever something is passed *by value* (as in `consume`), Rust interprets this as *ownership transfer*; when something is passed *by reference* (as in `add_something`), Rust interprets this as borrowing for a certain lifetime.

Shared references. Following the principle that we can have either aliasing or mutation, but not both at the same time, mutable references are unique pointers: they do not permit aliasing. To complete this picture, Rust has a second kind of reference, the *shared reference* written `&Vec<i32>` or `&'a Vec<i32>`, which allows aliasing but no mutation. One primary use-case for shared references is to share read-only data between multiple threads, as illustrated in Figure 5.

Here, we create a shared reference `vp_ptr` pointing to (and borrowing) `v[1]`. The vertical bars here represent a *closure* (also sometimes called an anonymous function or “lambda”) that does not take any arguments. These closures are passed to `join`, which is the Rust version of “parallel composition:” it takes two closures, runs both of them in parallel, waits until both are done, and returns both of their results. When `join` returns, the borrow ends, so we can mutate `v` again.

Just like mutable references, shared references have a lifetime. Under the

hood, the Rust compiler is using a lifetime to track the period during which `v` is temporarily shared between the two threads; after that lifetime is over (on line 5), the original owner of `v` regains full control. The key difference here is that multiple shared references are allowed to coexist during the same lifetime, so long as they are only used for *reading*, not writing. We can witness the enforcement of this restriction by changing one of the two threads in our example to `|| v.push(12)`: then the compiler complains that we cannot have a mutable reference and a shared reference to the `Vec` at the same time. And indeed, that program has a fatal data race between the reading thread and the thread that pushes to the vector, so it is important the compiler detects such cases statically.

Shared references are also useful in sequential code; for example, while doing a shared iteration over a vector we can still pass a shared reference to the *entire* vector to another function. But for this article, we will focus on the use of sharing for concurrency.

Summary. In order to obtain safety, the Rust type system enforces the discipline that a reference is never both aliased and mutable. Having a value of type `T` means you “own” it fully. The value of type `T` can be “borrowed” using a mutable reference (`&mut T`) or shared reference (`&T`).

Relaxing Rust’s Strict Ownership Discipline via Safe APIs

Rust’s core ownership discipline is sufficiently flexible to account for many low-level programming idioms.

But for implementing certain data structures, it can be overly restrictive. For example, without any mutation of aliased state, it is not possible to implement a doubly linked list because each node is aliased by both its next and previous neighbors.

Rust adopts a somewhat unusual approach to this problem. Rather than either complicating its type system to account for data structure implementations that do not adhere to it, or introducing dynamic checks to enforce safety at run time, Rust allows its ownership discipline to be relaxed through the development of *safe APIs*—APIs that extend the expressive power of the language by enabling safely controlled usage of aliased mutable state. Although the implementations of these APIs do not adhere to Rust’s strict ownership discipline (a point we return to later), the APIs themselves make critical use of Rust’s ownership and borrowing mechanisms to ensure that they preserve the safety guarantees of Rust as a whole. Let us now look at a few examples.

Shared mutable state. Rust’s shared references permit multiple threads to read shared data concurrently. But threads that just *read* data are only half the story, so next we will look at how the `Mutex` API enables one to safely share *mutable* state across thread boundaries. At first, this might seem to contradict everything we said so far about the safety of Rust: isn’t the whole point of Rust’s ownership discipline that it *prevents* mutation of shared state? Indeed it is, but we will see how, using `Mutex`, such mutation can be sufficiently re-

Figure 4. Use-after-free example with inferred reference lifetime.

```

1 let mut v = vec![10, 11];
2 let vp_ptr : &'a mut i32 = &mut v[1];
3 v.push(12);
4 println!("{}", *vp_ptr); // Compiler error
                               Lifetime 'a
    
```

Figure 5. Rust example: Shared references.

```

1 let v = vec![10,11];
2 let vp_ptr = &v[1];
3 join( || println!("v[1] = {}", *vp_ptr),
4      || println!("v[1] = {}", *vp_ptr));
5 v.push(12);
    
```

stricted so as to not break memory or thread safety. Consider the example in Figure 6.

We again use structured concurrency and shared references, but now we wrap the vector in a `Mutex`: the variable `mutex_v` has type `Mutex<Vec<i32>>`. The key operation on a `Mutex` is `lock`, which blocks until it can acquire the exclusive lock. The lock implicitly gets released by `v`'s destructor when the variable goes out of scope. Ultimately, this program prints either `[10,11,12]` if the first thread manages to acquire the lock first, or `[10, 11]` if the second thread does.

In order to understand how our example program type-checks, let us take a closer look at `lock`. It (almost^a) has type `fn(&'a Mutex<T>) -> MutexGuard<'a, T>`. This type says that `lock` can be called with a shared reference to a `mutex`, which is why Rust lets us call `lock` on both threads: both closures capture an `&Mutex<Vec<i32>>`, and as with the `vptr` of type `&i32` that got captured in our first concurrency example, both threads can then use that reference concurrently. In fact, it is crucial that `lock` take a shared rather than a mutable reference—otherwise, two threads could not attempt to acquire the lock at the same time and there would be no need for a lock in the first place.

The return type of `lock`, namely `MutexGuard<'a, T>`, is basically the same as `&'a mut T`: it grants exclusive access to the `T` that is stored inside the

a The actual type of `lock` wraps the result in a `LockResult<...>` for error handling, which explains why we use `unwrap` on lines 3 and 5.

lock. Moreover, when it goes out of scope, it automatically releases the lock (an idiom known in the C++ world as RAI³¹).

In our example, this means that both threads *temporarily* have exclusive access to the vector, and they have a mutable reference that reflects that fact—but thanks to the lock properly implementing mutual exclusion, they will never both have a mutable reference *at the same time*, so the uniqueness property of mutable references is maintained. In other words, `Mutex` can offer mutation of aliased state safely because it implements run-time checks ensuring that, *during* each mutation, the state is *not* aliased.

Reference counting. We have seen that shared references provide a way to share data between different parties in a program. However, shared references come with a *statically determined* lifetime, and when that lifetime is over, the data is uniquely owned again. This works well with structured parallelism (like `join` in the previous example) but does not work with *unstructured* parallelism where threads are spawned off and keep running independently from the parent thread.

In Rust, the typical way to share data in such a situation is to use an *atomically reference-counted* pointer: `Arc<T>` is a pointer to `T`, but it also counts how many such pointers exist and deallocates the `T` (and releases its associated resources) when the last pointer is destroyed. (This can be viewed as a form of lightweight library-implemented garbage collection.) Since the data is shared, we cannot obtain an `&mut T`

from an `Arc<T>`—but we *can* obtain an `&T` (where the compiler ensures that during the lifetime of the reference, the `Arc<T>` does not get destroyed), as in the example in Figure 7.

We start by creating an `Arc` that points to our usual vector. `arc_v2` is obtained by *cloning* `arc_v1`, which means that the reference count gets bumped up by one, but the data itself is not duplicated. Then we spawn a thread that uses `arc_v2`; this thread keeps running in the background even when the function we are writing here returns. Because this is unstructured parallelism we have to explicitly move (that is, transfer ownership of) `arc_v2` into the closure that runs in the other thread. `Arc` is a “smart pointer” (similar to `shared_ptr` in C++), so we can work with it almost as if it were an `&Vec<i32>`. In particular, in lines 3 and 4 we can use indexing to print the element at position 1. Implicitly, as `arc_v1` and `arc_v2` go out of scope, their destructors get called, and the last `Arc` to be destroyed deallocates the vector.

Thread safety. There is one last type that we would like to talk about in this brief introduction to Rust: `Rc<T>` is a reference-counted type very similar to `Arc<T>`, but with the key distinction that `Arc<T>` uses an *atomic* (fetch-and-add) instruction to update the reference count, whereas `Rc<T>` uses *non-atomic* memory operations. As a result, `Rc<T>` is potentially faster, but *not* thread-safe. The type `Rc<T>` is useful in complex sequential code where the static scoping enforced by shared references is not flexible enough, or where one cannot statically predict when the last reference to an object will be destroyed so that the object itself can be deallocated.

Since `Rc<T>` is not thread-safe, we need to make sure that the programmer does not accidentally use `Rc<T>` when they should have used `Arc<T>`. This is important: if we take our previous `Arc` example, and replace all the `Arc` by `Rc`, the program has a data race and might deallocate the memory too early or not at all. However, quite remarkably, the Rust compiler is able to catch this mistake. The way this works is that Rust employs something called the *Send trait*: a property of types, which is only enjoyed by a type `T` if elements of type `T` can be

Figure 6. Rust example: Shared mutable state.

```
1 let mutex_v = Mutex::new(vec![10, 11]);
2 join(
3   || { let mut v = mutex_v.lock().unwrap();
4       v.push(12); },
5   || { let v = mutex_v.lock().unwrap();
6       println!("{:?}", *v) });
```

Figure 7. Rust example: Reference counting.

```
1 let arc_v1 = Arc::new(vec![10, 11]);
2 let arc_v2 = Arc::clone(&arc_v1);
3 spawn(move || println!("{:?}", arc_v2[1]));
4 println!("{:?}", arc_v1[1]);
```


safely sent to another thread. The type `Arc<T>` is `Send`, but `Rc<T>` is not. Both `join` and `spawn` require everything captured by the closure(s) they run to be `Send`, so if we capture a value of the non-`Send` type `Rc<T>` in a closure, compilation will fail.

Rust's use of the `Send` trait demonstrates how sometimes the restrictions imposed by strong static typing can lead to *greater* expressive power, not less. In particular, C++'s smart reference-counted pointer, `std::shared_ptr`, always uses atomic instructions,^b because having a more efficient non-thread-safe variant like `Rc` is considered too risky. In contrast, Rust's `Send` trait allows one to “hack without fear.”²⁶ it provides a way to have both thread-safe data structures (such as `Arc`) and non-thread-safe data structures (such as `Rc`) in the same language, while ensuring modularly that the two do not get used in incorrect ways.

Unsafe Code, Safely Encapsulated

We have seen how types like `Arc` and `Mutex` let Rust programs safely use features such as reference counting and shared mutable state. However, there is a catch: *those types cannot actually be implemented in Rust*. Or, rather, they cannot be implemented in *safe* Rust: the compiler would reject an implementation of `Arc` for potentially violating the aliasing discipline. In fact, it would even reject the implementation of `Vec` for accessing potentially uninitialized memory. For efficiency reasons, `Vec` manually manages the underlying buffer and tracks which parts of it are initialized. Of course, the implementation of `Arc` does *not* in fact violate the aliasing discipline, and `Vec` does *not* in fact access uninitialized memory, but the arguments needed to establish those facts are too subtle for the Rust compiler to infer.

To solve this problem, Rust has an “escape hatch:” Rust consists not only of the safe language we discussed so far—it also provides some *unsafe* features such as C-style unrestricted pointers. The safety (memory safety and/or thread safety) of these features cannot be guar-

^b More precisely, on Linux it uses atomic instructions if the program uses `pthread`, that is, if it or any library it uses *might* spawn a thread.



We hope to inspire other members of the computer science research community to start paying closer attention to Rust and to help contribute to the development of this groundbreaking language.



anteed by the compiler, so they are only available inside syntactic blocks that are marked with the `unsafe` keyword. This way, one can be sure to not *accidentally* leave the realm of safe Rust.

For example, the implementation of `Arc` uses unsafe code to implement a pattern that would not be expressible in safe Rust: sharing without a clear owner, managed by thread-safe reference counting. This is further complicated by support for “weak references:” references that do not keep the referent alive, but can be atomically checked for liveness and upgraded to a full `Arc`. The correctness of `Arc` relies on rather subtle concurrent reasoning, and the Rust compiler simply has no way to verify statically that deallocating the memory when the reference count reaches zero is in fact safe.

Alternatives to unsafe blocks. One could turn things like `Arc` or `Vec` into language primitives. For example, Python and Swift have built-in reference counting, and Python has `list` as a built-in equivalent to `Vec`. However, these language features are implemented in C or C++, so they are not actually any safer than the unsafe Rust implementation. Beyond that, restricting unsafe operations to implementations of language primitives also severely restricts flexibility. For example, Firefox uses a Rust library implementing a variant of `Arc` without support for weak references, which improves space usage and performance for code that does not need them. Should the language provide primitives for every conceivable spot in the design space of any built-in type?

Another option to avoid unsafe code is to make the type system expressive enough to actually be able to verify safety of types like `Arc`. However, due to how subtle correctness of such data structures can be (and indeed `Arc` and simplified variants of it have been used as a major case-study in several recent formal verification papers^{9,12,18}), this basically requires a form of general-purpose theorem prover—and a researcher with enough background to use it. The theorem proving community is quite far away from enabling developers to carry out such proofs themselves.

Safe abstractions. Rust has instead opted to allow programmers the flexibility of writing unsafe code when nec-

essary, albeit with the expectation that it should be *encapsulated by safe APIs*. Safe encapsulation means that, regardless of the fact that Rust APIs like `Arc` or `Vec` are implemented with unsafe code, *users* of those APIs should not be affected: so long as users write well-typed code in the safe fragment of Rust, they should never be able to observe anomalous behaviors due to the use of unsafe code in the APIs' implementation. This is in marked contrast to C++, whose weak type system lacks the ability to even enforce that APIs are *used* safely. As a result, C++ APIs like `shared_ptr` or `vector` are prone to misuse, leading to reference-counting bugs and iterator invalidation, which do not arise in Rust.

The ability to write unsafe code is like a lever that Rust programmers use to make the type system more useful without turning it into a theorem prover, and indeed we believe this to be a key ingredient to Rust's success. The Rust community is developing an entire ecosystem of safely usable high-performance libraries, enabling programmers to build safe and efficient applications on top of them.

But of course, there is no free lunch: it is up to the author of a Rust library to somehow ensure that, if they write unsafe code, they are being very careful not to break Rust's safety guarantees. On the one hand, this is a much better situation than in C/C++, because the vast majority of Rust code is written in the safe fragment of the language, so Rust's "attack surface" is much smaller. On the other hand, when unsafe code is needed, it is far from obvious how a programmer is supposed to know if they are being "careful" enough.

To maintain confidence in the safety of the Rust ecosystem, we therefore really want to have a way of formally specifying and verifying what it means for uses of unsafe code to be safely encapsulated behind a safe API. This is precisely the goal of the **RustBelt** project.

RustBelt: Securing the Foundations of Rust

The key challenge in verifying Rust's safety claims is accounting for the interaction between safe and unsafe code. To see why this is challenging, let us briefly take a look at the standard technique for verifying safety of pro-

gramming languages—the so called *syntactic approach*.^{14,37} Using that technique, safety is expressed in terms of a *syntactic typing judgment*, which gives a formal account of the type checker in terms of a number of mathematical inference rules.

Theorem 1 (Syntactic type soundness). *If a program e is well-typed with respect to the syntactic typing judgment, then e is safe.*

Unfortunately, this theorem is too weak for our purposes, because it only talks about *syntactically* safe programs, thus ruling out programs that use unsafe code. For example, `if true { e } else { crash() }` is not syntactically well-typed, but it is still safe since `crash()` is never executed.

The key solution: Semantic type soundness. To account for the interaction between safe and unsafe code, we instead use a technique called *semantic type soundness*, which expresses safety in terms of the "behavior" of the program rather than a fixed set of inference rules. The key ingredient of semantic soundness is a *logical relation*, which assigns a *safety contract* to each API. It expresses that if the inputs to each method in the API conform to their specified types, then so do the outputs. Using techniques from formal verification, one can then prove that an implementation of the API satisfies the assigned safety contract, as depicted in Figure 8.

Semantic type soundness is ideal for reasoning about programs that use a combination of safe and unsafe code. For any library that uses unsafe code (such as `Arc`, `Mutex`, `Rc`, and `Vec`) one has to prove by hand that the implementation satisfies the safety contract. For example:

Theorem 2. *Arc satisfies its safety contract.*

For safe pieces of a program, the verification is automatic. This is expressed by the following theorem, which says that if a component is written in the

safe fragment of Rust, it satisfies its safety contract by construction.

Theorem 3 (Fundamental theorem). *If a component e is syntactically well-typed, then e satisfies its safety contract.*

Together, these imply that a Rust program is safe if the only appearances of unsafe blocks are within libraries that have been manually verified to satisfy their safety contracts.

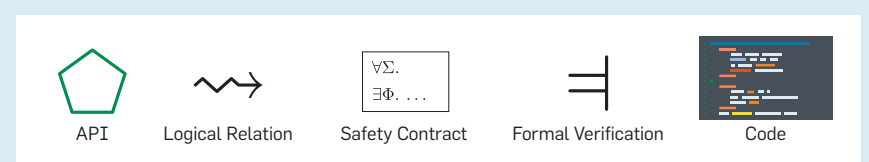
Using the Iris logic to encode safety contracts. Semantic type soundness is an old technique, dating back at least to Milner's seminal 1978 paper on type soundness,²⁸ but scaling it up to realistic modern languages like Rust has proven a difficult challenge. In fact, scaling it up to languages with mutable state and higher-order functions remained an open problem until the development of "step-indexed Kripke logical relations" (SKLR) models^{3,5} as part of the Foundational Proof-Carrying Code project^{2,4} in the early 2000s. Even then, verifications of safety contracts that were encoded directly using SKLR models turned out to be very tedious, low-level, and difficult to maintain.

In RustBelt we build upon more recent work on Iris,^{19–21,23} a verification framework for higher-order, concurrent, imperative programs, implemented in the Coq proof assistant.¹ Iris provides a much higher-level language for encoding and working with SKLR models, thus enabling us to scale such models to handle a language as sophisticated as Rust. In particular, Iris is based on *separation logic*,^{29,30} an extension of Hoare logic¹⁵ geared specifically toward modular reasoning about pointer-manipulating programs and centered around the concept of ownership. This provides us with an ideal language in which to model the semantics of ownership types in Rust.

Iris extends traditional separation logic with several additional features that are crucial for modeling Rust:

- Iris supports *user-defined ghost state*: the ability to define custom logi-

Figure 8. Semantic type soundness.



cal resources that are useful for proving correctness of a program but do not correspond directly to anything in its physical state. Iris’s user-defined ghost state has enabled us to verify the soundness of libraries like `Arc`, for which ownership does not correspond to physical ownership (for example, two separately owned `Arc<T>`’s may be backed by the same underlying memory)—a phenomenon known as “fictional separation.”^{10,11} It has also enabled us to reason about Rust’s borrowing and lifetimes at a much higher level of abstraction, by deriving (within Iris) a new, domain-specific “lifetime logic.”

► Iris supports *impredicative invariants*: invariants on the program state that may refer cyclically to the existence of other invariants.³² Impredicative invariants play an essential role in modeling central type system features such as recursive types and closures.

The complexity of Rust demands that our semantic soundness proofs be *machine-checked*, as it would be too tedious and error-prone to do proofs by hand. Fortunately, Iris comes with a rich set of *separation-logic tactics*, which are patterned after standard Coq tactics and thus make it possible to interactively develop machine-checked semantic soundness proofs in a time-tested style familiar to Coq users.^{22,24}

Conclusion and Outlook

In this article we have given a bird’s-eye view of Rust, demonstrating its core concepts like borrowing, lifetimes, and unsafe code encapsulated inside safe APIs. These features have helped Rust become the first industry-supported language to overcome the trade-off between safety and control.

To formally investigate Rust’s safety claims, we described the proof technique of semantic type soundness, which has enabled us to begin building a rigorous foundation for Rust in the RustBelt project. For more details about Rust and RustBelt, we refer the interested reader to our POPL’18 paper¹⁸ and the first author’s Ph.D. thesis.¹⁷

There is still much work left to do. Although RustBelt has recently been extended to account for the relaxed-memory concurrency model that Rust inherits from C++,⁹ there are a number of other Rust features and APIs that it does not yet

cover, such as its “trait” system, which is complex enough to have been the source of subtle soundness bugs.⁷ Moreover, although verifying the soundness of an internally unsafe Rust library requires, at present, a deep background in formal semantics, we hope to eventually develop formal methods that can be put directly in the hands of programmers.

Finally, while RustBelt has focused on building foundations for Rust itself, we are pleased to see other research projects (notably Prusti⁶ and RustHorn²⁷) beginning to explore an exciting, orthogonal direction: namely, the potential for Rust’s strong type system to serve as a powerful tool in simplifying the formal verification of systems code.

Acknowledgments

We wish to thank the Rust community in general, and Aaron Turon and Niko Matsakis in particular, for their feedback and countless helpful discussions. This research was supported in part by a European Research Council (ERC) Consolidator Grant for the project “RustBelt,” funded under the European Union’s Horizon 2020 Framework Programme (grant agreement no. 683289), and by the Dutch Research Council (NWO), project 016.Veni.192.259. ■

References

1. The Coq proof assistant, 2019; <https://coq.inria.fr/>.
2. Ahmed, A., Appel, A.W., Richards, C.D., Swadi, K.N., Tan, G. and Wang, D.C. Semantic foundations for typed assembly languages. *TOPLAS* 32, 3 (2010).
3. Ahmed, A.J. Semantics of types for mutable state. Ph.D. thesis, Princeton University, 2004.
4. Appel, A.W. Foundational proof-carrying code. *LICS*, 2001.
5. Appel, A.W. and McAllester, D. An indexed model of recursive types for foundational proof-carrying code. *TOPLAS* 23, 5 (2001).
6. Astrauskas, V., Müller, P., Poli, F. and Summers, A.J. Leveraging Rust types for modular specification and verification. *PACMPL* 3 (OOPSLA), 2019.
7. Ben-Yehuda, A. Coherence can be bypassed by an indirect impl for a trait object, 2019; <https://github.com/rust-lang/rust/issues/57893>.
8. Burch, A. Using Rust in Windows. Blog post, 2019; <https://msrc-blog.microsoft.com/2019/11/07/using-rust-in-windows/>.
9. Dang, H.-H., Jourdan, J.-H., Kaiser, J.-O. and Dreyer, D. RustBelt meets relaxed memory. *PACMPL* 4 (POPL), 2020.
10. Dinsdale-Young, T., Dodds, M., Gardner, P., Parkinson, M.J. and Vafeiadis, V. Concurrent abstract predicates. *ECCOP*, 2010.
11. Dinsdale-Young, T., Gardner, P. and Wheelhouse, M.J. Abstraction and refinement for local reasoning. *VSTTE*, 2010.
12. Doko, M. and Vafeiadis, V. Tackling real-life relaxed concurrency with FSL++. *ESOP 10201, LNCS*, 2017.
13. Grossman, D., Morrisett, G., Jim, T., Hicks, M., Wang, Y. and Cheney, J. Region-based memory management in Cyclone. *PLDI*, 2002.
14. Harper, R. *Practical Foundations for Programming Languages* (2nd Ed.). Cambridge University Press, 2016.
15. Hoare, C.A.R. An axiomatic basis for computer programming. *Commun. ACM* 12, 10 (1969).
16. Hofelt, D. Implications of rewriting a browser

- component in Rust. Blog post, 2019; <https://hacks.mozilla.org/2019/02/rewriting-a-browser-component-in-rust/>.
17. Jung, R. Understanding and Evolving the Rust Programming Language. Ph.D. thesis, Universität des Saarlandes, 2020; <https://people.mpi-sws.org/~jung/thesis.html>.
18. Jung, R., Jourdan, J.-H., Krebbers, R. and Dreyer, D. RustBelt: Securing the foundations of the Rust programming language. *PACMPL* 2 (POPL), 2018.
19. Jung, R., Krebbers, R., Birkedal, L. and Dreyer, D. Higher-order ghost state. *ICFP*, 2016.
20. Jung, R., Krebbers, R., Jourdan, J.-H., Bizjak, A., Birkedal, L. and Dreyer, D. Iris from the ground up: A modular foundation for higher-order concurrent separation logic. *JFP* 28 (2018).
21. Jung, R., Swasey, D., Sieczkowski, F., Svendsen, K., Turon, A., Birkedal, L. and Dreyer, D. Iris: Monoids and invariants as an orthogonal basis for concurrent reasoning. *POPL*, 2015.
22. Krebbers, R., Jourdan, J.-H., Jung, R., Tassarotti, J., Kaiser, J.-O., Timany, A., Charguéraud, A. and Dreyer, D. MoSeL: A general, extensible modal framework for interactive proofs in separation logic. *PACMPL* 2 (ICFP), 2018.
23. Krebbers, R., Jung, R., Bizjak, A., Jourdan, J., Dreyer, D. and Birkedal, L. The essence of higher-order concurrent separation logic. *ESOP*, 2017.
24. Krebbers, R., Timany, A. and Birkedal, L. Interactive proofs in higher-order concurrent separation logic. *POPL*, 2017.
25. Levick, R. Why Rust for safe systems programming. Blog post, 2019; <https://msrc-blog.microsoft.com/2019/07/22/why-rust-for-safe-systems-programming/>.
26. Matsakis, N. and Turon, A. Rust in 2016, 2015. Blog post; <https://blog.rust-lang.org/2015/08/14/Next-year.html>.
27. Matsushita, Y., Tsukada, T. and Kobayashi, N. RustHorn: CHC-based verification for Rust programs. *ESOP*, 2020.
28. Milner, R. A theory of type polymorphism in programming. *J. Computer and System Sciences* 17, 3 (1978).
29. O’Hearn, P.W., Reynolds, J.C. and Yang, H. Local reasoning about programs that alter data structures. *CSL*, 2001.
30. O’Hearn, P.W. Resources, concurrency, and local reasoning. *Theoretical Computer Science* 375, 1-3 (2007).
31. Stroustrup, B. *The C++ Programming Language*. Addison-Wesley, 2013.
32. Svendsen, K. and Birkedal, L. Impredicative concurrent abstract predicates. *ESOP*, 2014.
33. Thomas, G. A proactive approach to more secure code. Blog post, 2019; <https://msrc-blog.microsoft.com/2019/07/16/a-proactive-approach-to-more-secure-code/>.
34. Tofte, M. and Talpin, J. Region-based memory management. *Information and Computation* 132, 2 (1997).
35. Tu, T., Liu, X., Song, L. and Zhang, Y. Understanding real-world concurrency bugs in Go. *ASPLOS*, 2019.
36. Walker, D. Substructural type systems. *Advanced Topics in Types and Programming Languages*. B.C. Pierce, Ed. MIT Press, Cambridge, MA, 2005.
37. Wright, A.K. and Felleisen, M. A syntactic approach to type soundness. *Information and Computation* 115, 1 (1994).

Ralf Jung is a postdoc at the Max Planck Institute for Software Systems, Germany.

Jacques-Henri Jourdan is a researcher at the Université Paris-Saclay, CNRS, ENS Paris-Saclay, Laboratoire des méthodes formelles, France.

Robbert Krebbers is a (tenured) assistant professor at Radboud University Nijmegen, The Netherlands.

Derek Dreyer is a professor at the Max Planck Institute for Software Systems, Germany.



Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/programming-in-rust>

Attention: Undergraduate and Graduate Computing Students

There's an **ACM Student Research Competition (SRC)**
at a SIG Conference of interest to you!



Association for Computing Machinery
Advancing Computing as a Science & Profession



It's hard to put the **ACM Student Research Competition** experience into words, but we'll try...



"Attending ACM SRC was a transformative experience for me. It was an opportunity to take my research to a new level, beyond the network of my home university. Most important, it was a chance to make new connections and encounter new ideas that had a lasting impact on my academic life. I can't recommend ACM SRC enough to any student who is looking to expand the horizons of their research endeavors."

David Mueller
North Carolina State University | SIGDOC 2018



"Participating in the ACM SRC was a unique opportunity for practicing my presentation skills, getting feedback on my work, and networking with both leading researchers and fellow SRC participants. Winning the competition was a great honor, a motivation to continue working in research, and a useful boost for my career. I highly recommend any aspiring student researcher to participate in the SRC."

Manuel Rigger
Johannes Kepler University Linz, Austria | Programming 2018



"The SRC was a great chance to present early results of my work to an international audience. Especially the feedback during the poster session helped me to steer my work in the right direction and gave me a huge motivation boost. Together with the connections and friendships I made, I found the SRC to be a positive experience."

Matthias Springer
Tokyo Institute of Technology | SPLASH 2018



"I have been a part of many conferences before both as an author and as a volunteer but I found SRC to be an incredible conference experience. It gave me the opportunity to have the most immersive experience, improving my skills as a presenter, researcher, and scientist. Over the several phases of ACM SRC, I had the opportunity to present my work both formally (as a research talk and research paper) and informally (in poster or demonstration session). Having talked to a diverse range of researchers, I believe my work has much broader visibility now and I was able to get deep insights and feedback on my future projects. ACM SRC played a critical role in facilitating my research, giving me the most productive conference experience."

Muhammad Ali Gulzar
University of California, Los Angeles | ICSE 2018



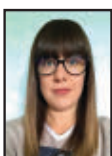
"At the ACM SRC, I got to learn about the work done in a variety of different research areas and experience the energy and enthusiasm of everyone involved. I was extremely inspired by my fellow competitors and was happy to discover better ways of explaining my own work to others. I would like to specifically encourage undergraduate students to not hesitate and apply! Thank you to all those who make this competition possible for students like me."

Elizaveta Tremsina
UC Berkeley | TAPIA 2018



"The ACM SRC was an incredible opportunity for me to present my research to a wide audience of experts. I received invaluable, supportive feedback about my research and presentation style, and I am sure that the lessons I learned from the experience will stay with me for the rest of my career as a researcher. Participating in the SRC has also made me feel much more comfortable speaking to other researchers in my field, both about my work as well as projects I am not involved in. I would strongly recommend students interested in research to apply to an ACM SRC—there's really no reason not to!"

Justin Lubin
University of Chicago | SPLASH 2018



"Joining the Student Research Competition of ACM gave me the opportunity to measure my skills as a researcher and to carry out a preliminary study by myself. Moreover, I believe that "healthy competition" is always challenging in order to improve yourself. I suggest that every Ph.D. student try this experience."

Gemma Catolino
University of Salerno | MobileSoft 2018

Check the SRC Submission Dates: <https://src.acm.org/submissions>

- ◆ Participants receive: \$500 (USD) travel expenses
- ◆ All Winners receive a medal and monetary award. First place winners advance to the SRC Grand Finals
- ◆ Grand Finals Winners receive a handsome certificate and monetary award at the ACM Awards Banquet

Questions? Contact Nanette Hernandez, ACM's SRC Coordinator: hernandez@hq.acm.org



Attention, particularly self-attention, is a standard in current NLP literature, but to achieve meaningful models, attention is not enough.

BY EDUARDO SOUZA DOS REIS, CRISTIANO ANDRÉ DA COSTA, DIÓRGENES EUGÊNIO DA SILVEIRA, RODRIGO SIMON BAVARESCO, RODRIGO DA ROSA RIGHI, JORGE LUIS VICTÓRIA BARBOSA, RODOLFO STOFFEL ANTUNES, MÁRCIO MIGUEL GOMES, AND GUSTAVO FEDERIZZI

Transformers Aftermath: Current Research and Rising Trends

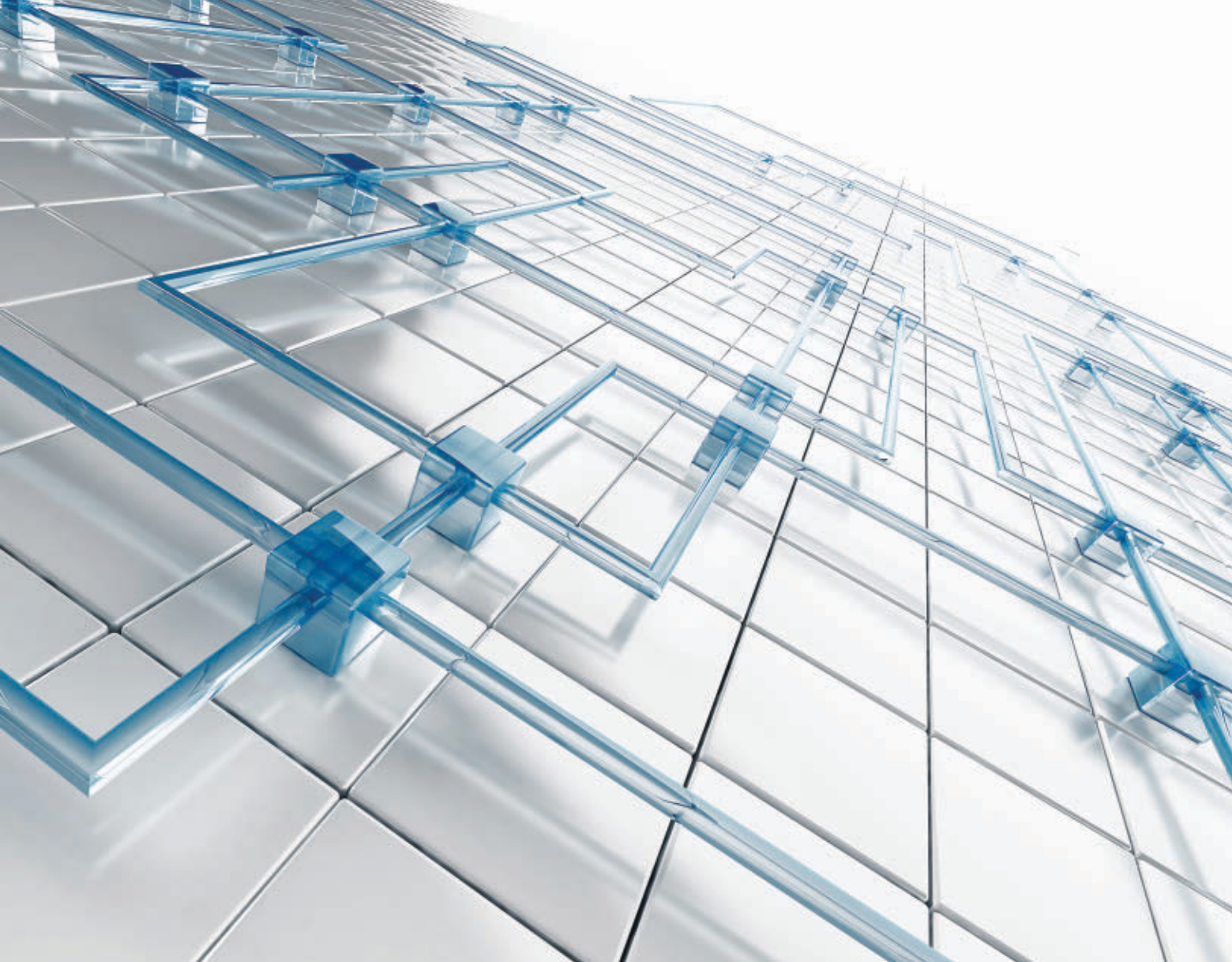
NATURAL LANGUAGE PROCESSING (NLP) is a main subject for artificial intelligence research, on par with computer vision and scene understanding. Knowledge is encoded and transferred among individuals through text, following a formally defined structure. Nonetheless, distinct languages, context, and subtle particularities of different communication channels are complex challenges that researchers must cope with. Hence, the task of general language modeling and understating was divided into multiple subtasks.

For example, question and answering, image captioning, text summarization, machine translation and natural language generation. Recently, attention mechanisms became ubiquitous among state-of-the-art approaches, allowing the models to selectively attend to different words or sentences in order of relevance. The goal of this review is to gather and analyze current research efforts that improve on—or provide alternatives to—attention mechanisms, categorize trends, and extrapolate possible research paths for future works.

Recurrent Neural Networks (RNNs) were broadly adopted by the NLP community and achieved important milestones.³⁷ However, it is computationally expensive to encode long-term relations among words in a sentence, or among sentences in a document using RNNs. In tasks such as text generation, encoding these dependencies is fundamental, and the inference time may become prohibitively slow. Pursuing a solution for these limitations, the seminal work of Vaswani et al.³³ engineered the Transformers architecture. The core idea was to put attention mechanisms in evidence, discarding recurrence. Soon after, the Transformers became the de facto backend model for most NLP. From the original December 2017 publication to date, there

» key insights

- **On current models, the produced text has surprisingly middling segments contrasting with meaningless word sequences. This lack of meaning cannot be trivially solved through attention alone.**
- **The state of the art is still heavily reliant on the original BERT as a backend model. They were focusing on either increasing the number of learnable parameters or their cost-effectiveness.**
- **The performance on most few-shot scenarios is way lower than human baselines. It further highlights the need for additional knowledge encoding methods, such as Knowledge Graphs. Consequently, few-shot learning is a vast research trend and will focus on the community for the years to come.**



have been over 4,000 citations and remarkable work on the concept. Due to its early success, the community focus became the development of larger models, containing scaled-up Transformers and longer context windows, leading to a shift from task-specific training procedures to more general language modeling. On the other hand, although current models are capable of generating text with unprecedented perplexity, they rely mostly on statistical knowledge contained in the text corpora, without encoding the actual meaning behind words. Hence, the produced text has surprisingly middling segments contrasting with meaningless word sequences. This lack of meaning cannot be trivially solved through attention alone.

Current Research

Most NLP tasks can be designed under a sequence-to-sequence modeling

framework. For instance, a sequence of image pixels being transformed in a caption composed of sequences of words; or a sentence in a given input language being translated to a correspondent sentence in an output language. On both examples, the input tokens (for example, pixels or words) must be mapped to the output tokens (words) by similarity. Sequence-to-sequence modelling is defined as mapping the distribution probability of the next token y_n on the output sequence $y = (y_1, y_2, \dots, y_{n-1})$, given a variable length input sequence $x = (x_1, x_2, \dots, x_n)$. In practice, the mapping is not done directly and the model also yields intermediate representations $z = (z_1, z_2, \dots, z_n)$, which, ideally, encode the context that the input tokens are exposed to.⁵ Such intermediate representations are referred to as context vectors in the NLP literature, and the two-step procedure as encoder-decoder methods.

Initially, the community leaned toward recurrent models for both decoders and encoders, mainly to the RNNs.⁸ Part of the reviewed works rely on sentence-level embeddings,²⁴ but here we assume that the set of input tokens is a vector of words encoded in a continuous representation.²² Further, for the sake of clarity, we will provide mostly examples based on machine translation, which was the main target for the original Transformers.

Recurrent Neural Networks. A RNN receives the input sequence x and processes it in a linear fashion by updating its hidden state h_i at every timestep i . The updating procedure follows $h_i = f(h_{i-1}, x_i)$, where f is an activation function that provides non-linearity⁴ and x_i is the received input token at timestep i . In short, regarding NLP tasks, RNNs learn to predict the next input x_{i+1} given the distribution of previous ones. Nonetheless, RNN-based models have

clear shortcomings. For one, relying on an unidirectional pipeline, in which the network has access to the full context at the end of a sentence, but earlier iterations do not have information on the incoming tokens. The solution, previous to the Transformers, was to use bidirectional models stacking multiple layers of RNNs, as in the well-known work of Bahdanau et al.² on attention-based methods for NLP. Yet, activations for earlier tokens had a tendency to fade out through the pipeline, and long-term relations were lost.³³ As an example, in the sentence “The student failed the class due to stress,” the token “stress” is correlated to the “student” and not to the “class,” but the influence of the token “student” on the hidden state may have vanished by the time the model reaches the “stress” token. In order to solve this short-memory constraint (vanishing gradients), RNNs enhanced by *forgetting mechanisms* (units that manipulate the hidden state non-sequentially) got in evidence. Most reviewed approaches were based on either LSTMs³¹ GRU.⁵ Unfortunately, for all the aforementioned recurrent models, the time to train increases exponentially as the size of the context vector increases. In

turn, constraining its size makes so that bigger context windows end up having worse representations, that is, the information must be compressed more aggressively.

From encoder-decoder to attention.

Alongside RNNs, recent sequence-to-sequence models were based on the encoder-decoder architecture.⁴ As a general overview of such methods, we provide a brief explanation. The encoder is an RNN, fed with the set of input tokens x . It encodes x in a context vector, represented by its hidden state at current timestep i , where $h_i \in (h_0, h_1, \dots, h_n)$. The context vector changes over time but keeps a fixed length throughout the whole encoding process, while the size of the input set is variable. Similarly, the decoder is also a RNN, but instead of a sequence of tokens, it is fed the last hidden state h_n from the encoder. It yields a set of output tokens (y_1, y_2, \dots, y_n) generated one by one, given $p(y|d_j, h_n)$, where d_j is the decoder’s hidden state at time step j . This standard pipeline is illustrated in Figure 1(a).

Currently, the correlation between context and output is augmented through attention. Within the attention framework, the encoder yields the full set of hidden state vectors at

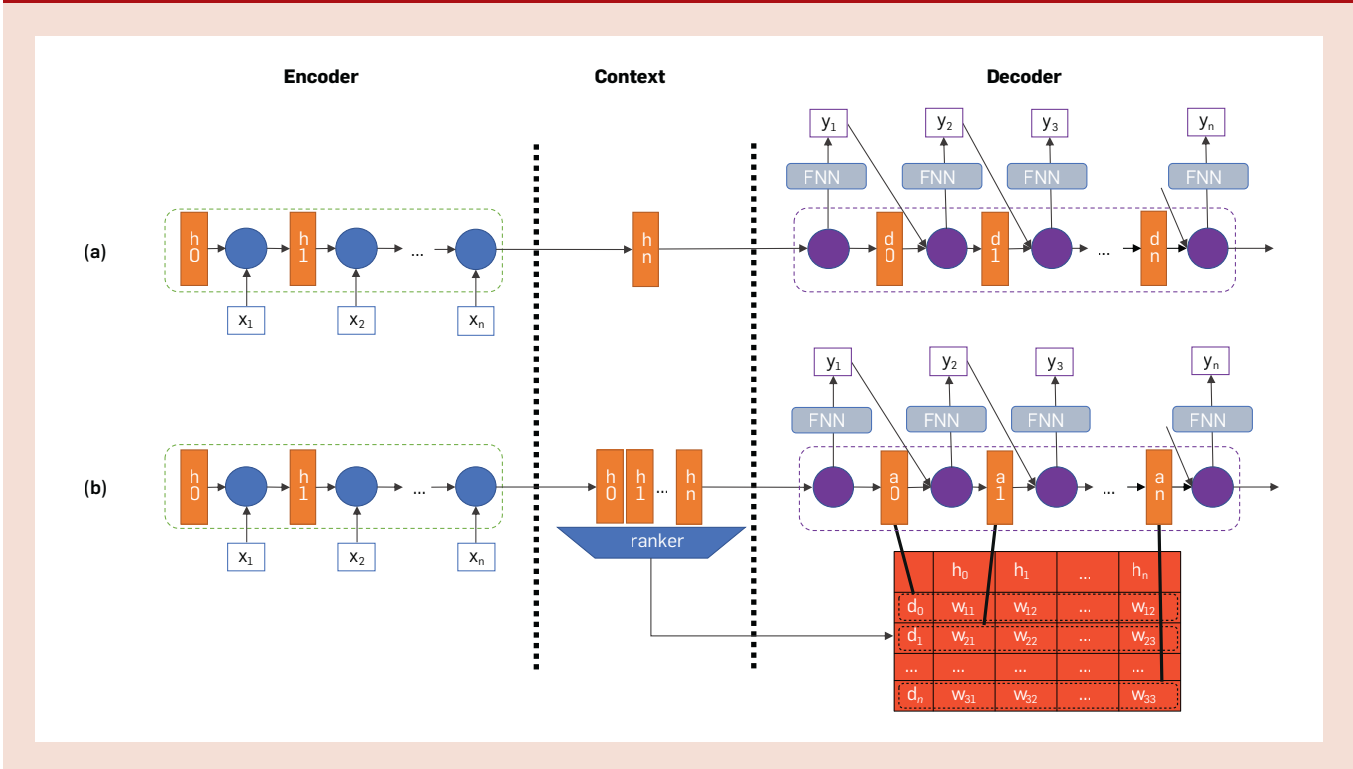
all iterations, instead of only the last one. Afterward, a classifier (ranker) is applied to the yielded set in order to measure the relevance of each hidden state with respect to each output token. Having a measure of relevance is the main aspect of attention mechanisms. More formally:

$$a_j = d_j (h_i \times l(w_{ij})) \tag{1}$$

where l is typically a softmax function, and $h_i \times l(w_{ij})$ is the weighted context vector for the decoder hidden state d_j . This enhanced pipeline is depicted in Figure 1(b).

Transformers. In 2017, Vaswani et al.³³ proposed the Transformers model arguing that there is no need for recurrence nor forgetting on sequence-to-sequence modeling. Transformers are a simpler architecture composed of a stack of encoders and a stack of decoders. Each encoder is an identical layer to other encoders, without weight sharing, and the same holds for decoders. In the context of machine translation, the topmost decoder uses a feedforward neural network (FNN) to provide logits, which feed a softmax layer to yield a probability distribution for the next output token.⁶

Figure 1. Difference between encoder-decoder methods (a) without and (b) with attention. Notice that the circles represent the same set of weights changing at different timesteps.



Formally, in the Transformers architecture, context is represented by the following simplified set of formulas, called *Scaled Dot-Product Attention*:

$$W(q_i, k) = q_i k^T \tag{2}$$

$$W'_i = \text{softmax}\left(\frac{W(q_i, k)}{\sqrt{d_k}}\right) \tag{3}$$

$$z_i = W'_i \times v \tag{4}$$

where the vector of inputs x is project into three new vector spaces: *query* (decoder), *keys* and *values* (encoder), represented by q, k and v . Keys can be seen as feature labels, while the values are its potentials regarding x . Queries define which features the previous decoder demanded. In addition, d_k stands for the dimensionality of the vector of keys. These projections are made with different trainable weight matrices, initialized randomly during training. In the first step, represented by Equation 2, the query from token i goes through the dot-product with k , containing all keys. Such operation yields the similarity of vector q_i with regard to each of the keys. Hence, the higher the dot-product, the more informative (relevant) the given key (feature) is to the proposed query. Next, through softmax, the method guarantees that the relevance will sum

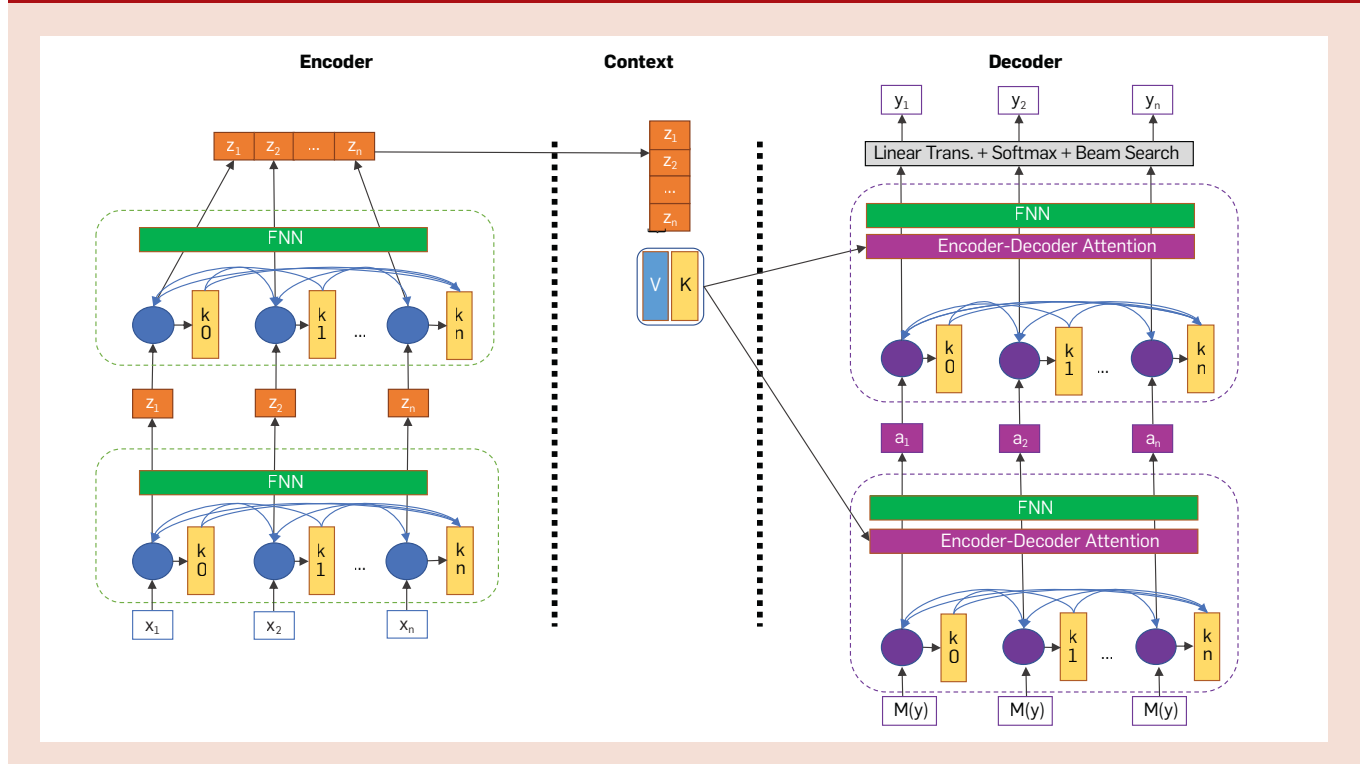
up to 1 and be positive, yielding a probability distribution over all keys with regards to q . Such distribution indexes the values vector.

In other words, the softmax use the exponential function to increase the gap between weights of relevant keys and less relevant ones, avoiding ambiguity. The resulting set of weights $W(q_i, k)$ goes through cross-product with the vector of values (Equation 4) to construct the context vector z , composed of weighted values.

Contrary to additive attention, which weights the values by the sum of q and k , multiplicative attention can be defined using matrices. Hence, it is faster to calculate in current hardware, but grows too fast as d_k gets larger, overflowing the softmax function. This is the reasoning behind the normalizing factor $1/d_k$ in Equation 2. Also, during training the decoder receives the target sentence as input. Future tokens are masked out, so it has no access to tokens that would be unavailable outside the training scope. For simplicity, we refer to this set of masked inputs as $M(y)$ and abstract secondary steps (residual connections and positional encoding). Figure 2 further illustrates the Transformers' main modules.

Advantages of attention mechanisms. Instead of increasing the length of the context vector, Transformers selectively look for the most informative tokens at each timestep, through multiple forms of attention. There is self-attention on both the encoder and the decoder, and there is a global encoder-decoder attention. Self-attention, is an attention mechanism that correlates different positions of the same sequence. Through self-attention, each cell of the context vector is informed by all previous inputs, resulting in a large receptive field over the whole sentence.⁸ It was a paradigm shift from RNNs, since it would backpropagate gradients for the whole sentence, instead of one input token at a time. Therefore, reducing the number of computational steps that information has to flow through in order to have an impact. Self-attention further provides a higher degree of interpretability. Interpretability is a measure of how understandable by a human the model and its predictions are.¹¹ Relying on black box models as the weights of a deep neural network. (DNN) lacks semantic interpretation. In contrast, some ablation studies on the original Transformers point out that the language model exhibits a behavior

Figure 2. An example of Transformers composed of two encoders and two decoders. Notice that the decoders receive the context—projected in two vectors v and k —from the topmost encoder.



related to both the syntactical and semantic structure of the sentences. Pushing the degree of interpretability of DNNs models is in itself a huge advantage in favor of attention. Additionally, the number of sequential operations increases linearly with the input. Thus, self-attention layers are faster than recurrent layers when the input sequence is shorter than the context vector, which happens quite often in practice.³³

The original Transformers also implements multihead attention: each attention head gives a different set of weights, similarly to an ensemble model. Assuming a sentence as input, one head would attend to the relation among subject and action, while others attend to adjectives and pronouns. Initially, the number of heads was chosen arbitrarily, yet, pruning and selection methods are currently available.³⁴ Since RNNs update their hidden state one input at a time sequentially, they are inherently slow to train. In contrast, Transformers have dependencies between inputs in the self-attention layer, but the FNN is shared by all context vectors and can therefore be run in parallel, as well as each of the attention heads.

Following, we give an overview of the main milestones achieved due to the Transformers architecture and its self-attention mechanism. Yet, some alternative methods were also impactful in the last couple years, such as the Conv Seq2Seq¹⁰ model. Comparing RNNs to Convolutional Networks (CNNs) for sequence-to-sequence modeling, CNNs are more parallelizable, and map the context into a layer-based hierarchical structure, in which long dependencies are naturally captured by higher layers. Besides the architecture, Gehring et al.¹⁰ also proposed an alternative attention method, the multi-step attention, in which the decoder receives a matrix of attention weights from the previous decoder and then calculate its own, instead of sharing the same attention weights matrix across all decoders. As an early example of a mixed approach, the Universal Transformers⁸ (UT) combined the parallelism of the Transformers, with the recurrent inductive bias of RNNs, which seems to be better suited to a range of sequence-to-sequence problems. At each recurrent

timestep, UT applies a self-attention mechanism and generates a context vector that attend to all input tokens. After, it applies a transition function to the next timestep, instead of the next encoder or decoder.

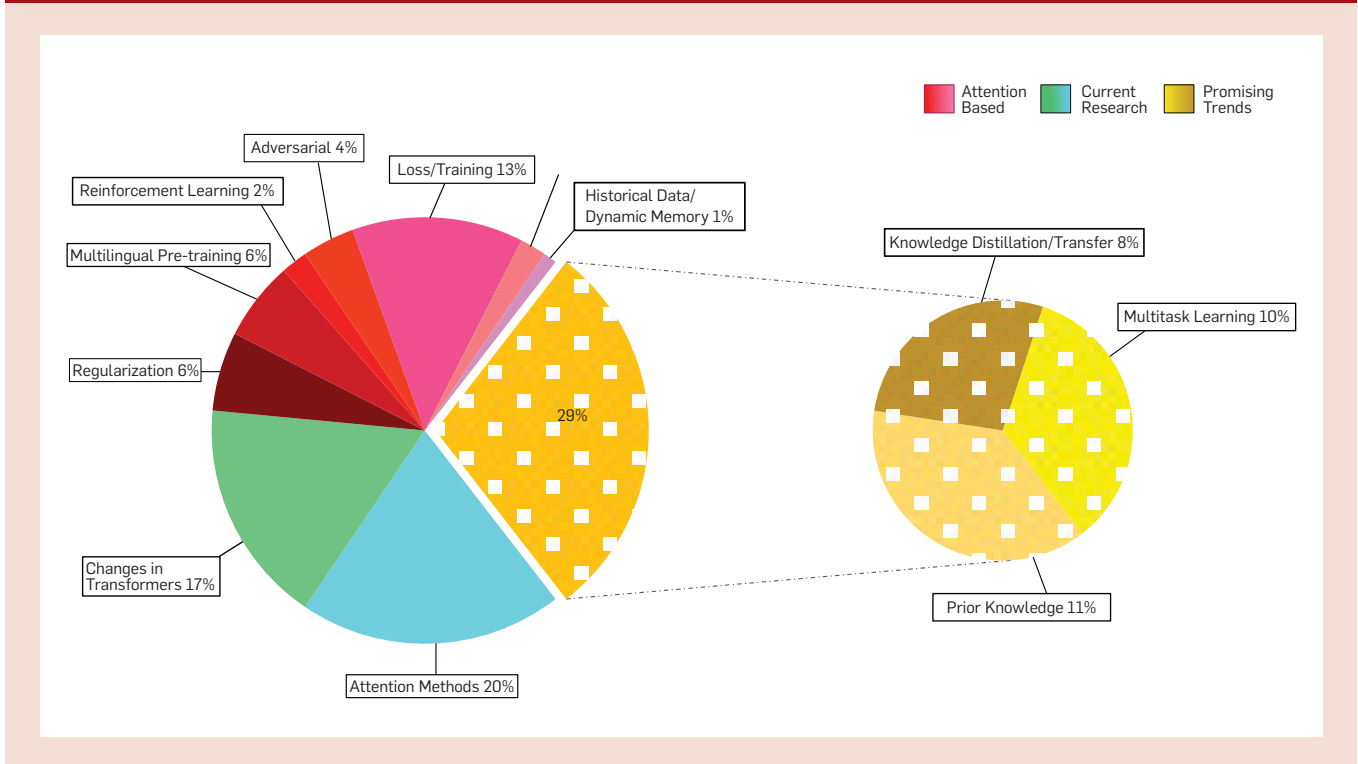
Unsupervised transformers. Language modeling is a key task for NLP, requiring the detection of very long-term dependencies. Given x as input, language modeling can be defined as the conditional probability $p(x_i) = p(x_i | (x_1, x_2, \dots, x_{i-1}), \theta)$, where θ is the set of model weights, and the context $(x_1, x_2, \dots, x_{i-1})$ is encoded as a vector z . Semi-supervised approaches use unsupervised pre-training—with language modeling as objective—and then fine-tune the model in a task-specific dataset. An advantage of unsupervised pre-training is that it implicitly adds a regularization step, enabling better generalization.²⁴ In addition, by removing the need of human labeling, it allows for the use of larger datasets. It was a huge shift in the recent NLP literature to transfer entire pretrained models among distinct works, instead of only word embeddings.²⁹

Two earlier examples of successful semi-supervised approaches are the ELMo²³ and the OpenAI Generative Pretrained Transformers²⁴ (GPT). In particular, the GPT is based entirely on the Transformers architecture. GPT's authors assume that since the objective is to predict the next token, the encoder stack can be discarded, and the decoders are directly correlated to the task. Contrary to other semi-supervised methods, the GPT model does not change its architecture for each sub-task, rather, it converts the input and output into an ordered sequence that the model can process, relying solely on the language model yielded by the training procedure. For example, on question and answering, the model concatenates the given document, the question and the set of answers, delimiting the input (document + question) by a \$, followed by the output (answers). Then, GPT predicts a probability distribution over all possible answers. It was the first task agnostic model able to outperform state-of-the-art methods.²⁴ Task agnostic refers to maintaining a single model architecture across all tasks. For example, the work of Radford et al.²⁵ used “TL;DR”

as a delimiter between input and target on summarization tasks.

Unidirectional language models such as the GPT provides narrow self-attention modules,⁹ that is, they apply a left-to-right architecture where the tokens can only attend to previous tokens. In contrast, another variation of the original Transformers named Bidirectional Encoder Representations from Transformers⁹ (BERT) used a bidirectional unsupervised pre-training. It hides parts of the input using a mask and forces the model to fill these gaps relying solely on context. Consequently, it allows for the current token to attend to both previous (left-to-right) and future (right-to-left) tokens. This bidirectional context-awareness has been explored for LSTMs as well, in the aforementioned work of Peters et al.²³ on ELMo. BERT is still present as part of state-of-the-art models to date. Yet, the modern BERT approach has changed overtime. In the original paper there are two losses: masked language modeling and next sentence prediction (NSP). NSP is a binary classification loss to predict if two sentences follow each other in a document, which has proven to be inefficient.³⁹ Nonetheless, modeling the relationship between sentences is an important aspect of language understanding.¹⁵ As an alternative to NSP, Lan et al.¹⁵ proposed sentence-order prediction loss, in which positive examples are two consecutive sequences in a document, just like in NSP, but there are also negative examples in inverse order, lacking coherence. Thus, the model is led into learning textual coherence features.

Radford et al.²⁵ argue that focusing on the development of larger task-specific datasets will be a hard path due to the scale to which the current models are conditioned; and the answer is to develop new unsupervised models through multitask learning. Methods that either fine-tune, or are solely trained, on task specific datasets are limited to the presented context, while completely unsupervised training leads to general architectures, conditioned to much richer language models. Recently, the authors proposed the GPT-2 model, arguing that Transformers are flexible enough to allow for task agnostic models given enough data. In order to justify such claims, they applied their

Figure 3. Impact of Transformers on the NLP literature.

language model to various tasks using zero-shot (no fine-tuning) and, quite remarkably, achieved state-of-the-art performance on some of them.

What Lies Beyond?

The objective of this review is to explore the recent approaches that use attention mechanisms to map long-term relations between tokens (words, sentences, documents) in NLP tasks. We applied a variation of the systematic review process,¹⁴ by hard constraining the set of articles to those that either cite the original Transformers, improve on benchmark results, or had a widely known contribution (number of citations) to the state of the art. After reviewing recent work, our final corpus^a comprises of 485 articles from 2018 and 2019, separated in 9 categories based on employed methods. Figure 3 details the proposed categorization of the reviewed literature. Although Multitask learning can be framed as regularization, we separate them from works that target other forms of regularization, since they are a main cluster among them. The same holds for Multilingual pretraining. Data

augmentation without new method proposals were discarded, due to the reliability on the state-of-the-art.

The review process converged on two main open issues that go beyond the capabilities of current attention mechanisms: commonsense reasoning and multitask learning. For the former, even GPT-2 language model yields plausible, yet, meaningless text in the sense that there is no understanding of words, just statistical knowledge on the distribution of these words in the training set. This statement can be further asserted by RNNs outperforming Transformers on simple tasks when the sentences' length during test differ too much from the ones used during training,⁸ or by the problem of non-literal meaning being completely missed by current solutions. To solve this limitation, we point out to the trend of using Knowledge Graphs (KGs) within unsupervised pretraining. The interpretability of KGs is explicit, while attention mechanisms may yield questionable explanations.¹³ In addition, KGs allow for transfer learning, and the use of smaller datasets, since part of the domain's features are previously encoded in the graph itself.

Another plausible alternative to the laborious task of developing larger

datasets is to explore data augmentation and increase data diversity instead of volume. For example, Yu et al.⁴⁰ applied a data augmentation technique for question and answering consisting of translating the context to other languages, providing another perspective to the same sentence (paraphrasing). However, data augmentation tends to be task-specific, and generating large synthetic datasets may be unfeasible.³²

With respect to multitask learning, to date, semi-supervised methods still outperform fully unsupervised ones given the same model size. Nonetheless, the scores achieved by GPT-2 on multiple NLP tasks point to unsupervised training being a core step toward multitask models. Authors of GPT-2 even speculate that, given sufficient capacity, the language model starts to infer the task itself. It is a reasonable speculation, since the global minimum for both supervised and unsupervised training procedures is the same. Supervision just constrains the search space. On the other hand, unsupervised scenarios are much slower to train, and have no guarantee of convergence. Albeit aforementioned achievements, GPT-2 performance is close to random on some tasks.

^a The full set of articles can be accessed at: <https://github.com/diorgenesugenio/transformers-aftermath>

We frame multitask learning as the closest milestone after unsupervised pretraining. As an example of applicability, BERT learns universal word representations that can be used for various tasks, yet, after fine-tuning this generalization is lost due to overfitting. Multitask learning is a plausible solution for such cases.¹⁷ Figure 4 provides an overview of a complete NLP model, based on promising research trends. Next, we detail the most successful solutions proposed so far: enhancing the original Transformers and exploring domain knowledge.

Enhancing the Transformers. The Transformer-XL⁶ model achieved superior results for language modeling than both the Transformers and RNNs, by being able to map longer dependencies among input tokens. The authors argue that the main limitation of original Transformers on context-heavy tasks is to encode all the context information on a fixed size vector. Just splitting the context itself may break important dependencies due to incorrect boundaries. Thus, the authors proposed to cache the current context representation, and propagate information from previous segments to further recurrent steps. Nonetheless, models combining recurrence and Transformers are a minority in the researched corpus, the state-of-the-art is heavily reliant on the original BERT as backend model.

BERT-based models. In their work on RoBERTa, Liu et al.²⁰ argue the original BERT was undertrained and could reach state-of-the-art performance by increasing training time alone. The main proposed changes to the training procedures were the use of bigger batches and removing the NSP step, both which became standards in the posterior literature. Specifically, larger batches as a way to improve training efficiency was first proposed, within the Transformers context, on You et al.³⁹ Their layer-wise adaptive large batch optimization allowed for batch sizes of 32868 sequences of tokens, achieving the memory limit of their test TPU. Remarkably, fully consuming the hardware reduced their training time from 3 days up to 76 minutes.

A multitask alternative was proposed in the work of Liu et al.¹⁸ on the Multi-Task DNN (MT-DNN). It is a BERT-based model, composed of a set of shared lay-

ers, and a set of task-specific sub-layers. The input sequence x is fed to the shared layers, and a bidirectional encoder captures contextual information through self-attention. Afterward, both contextual information and x are used as input to task-specific layers. BERT-based models can even be extrapolated beyond the scope of NLP, for instance, the ViLBERT model²¹ combined language modeling to visual inputs, aiming at applications such as image captioning. The authors propose a two-streams model, one for text and other for images, interacting through co-attention transformer layers. Co-attention refers to exchanging keys and values vectors among different attention heads (multi-head attention), mixing contextual features captured on both the language and the visual streams.

Pretraining. Among pretraining objectives, the two most successful³⁸ approaches are either Autoregressive (AR) language modeling or Autoencoding (AE), such as GPT and BERT, respectively. AR language modeling seeks to estimate the probability distribution of a text corpus, while AE reconstruct original data from partial input. The advantage of the latter is to enable bidirectional context encoding. In turn, artificially masking out input tokens results in a discrepancy between unsupervised pre-training and fine-tuning (no masks). Combining the two approaches is a recent trend that achieved new milestones in the works of Yang et al.³⁸ and Song et al.³⁰

Yang et al.³⁸ propose the XLNet to learn the dependency $\log p(x|U)$, where U is a subset of tokens in x that encodes its context. Differently from AR models, it can map dependencies of x and U regardless of the order. In other words, instead of adopting left-to-right or right-to-left models, it learns with respect to all possible permutations of tokens in x . XLNet outperformed the original BERT and other contemporary models on 18 out of 20 proposed tasks.

Song et al.³⁰ is a successful training procedure for encoder-decoder based natural language generation (NLG), which is a key application of language modeling, widely impacted by Transformers.²⁵ Their model is the MAsked Sequence to Sequence pretraining (MASS). Both BERT and GPT-2 train encoders and decoders separately, while

MASS trains both jointly. The encoder receives a subset of x with randomly masked sets of consecutive tokens. Next, the decoder receives the remaining subset masked and must learn the context only from the masked inputs, which are now available. Consequently, the encoder is forced to extract more context information in its hidden state to aid the decoder. Also, the decoder relies only on context, thus, it must be able of language understanding, that is, encoding meaning of words.

Larger models. A milestone on the trend of training ever larger models was achieved by NVIDIA in the Megatron-LM²⁹ paper. The authors noticed that precise placement of the normalization step was important on very large models. By manipulating residual connection, and the placement of normalization layers, their model showed monotonically increasing in performance as the model grew larger. For faster training, each attention head was processed in a different GPU, enabling a new level of parallelism. Due to the scalability of their model, they achieved state-of-the-art on GLUE tasks by increasing the size of the available BERT model from 340M up to 3.9B.

In contrast to the parallel nature of the Megatron-LM model, ALBERT¹⁵ provides a method to reduce the number of parameters, without an equivalent loss in performance. By improving the model scaling rather than its size, they are the current state-of-the-art on the GLUE³⁶ benchmark, while having fewer parameters than the original BERT. Two methods are applied: sharing parameters across all layers, enabling deeper networks without overflowing the number of parameters, and decomposing the word embedding matrix in smaller sets, enabling smaller hidden layers. Noteworthy, although ALBERT has less parameters than BERT, it is more computationally expensive due to its architecture.

Two models stand out in size. Turing-NLG^b is a 17 billion parameter language model by Microsoft., which follows a similar training procedure to Megatron-LM, but in a larger scale. Following the architecture proposed for

^b Available at <https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>

GPT-2, Brown et al.³ proposes GPT-3, a 175 billion parameters autoregressive model, which was a huge increase in size from the previous largest model in the literature. It was widely evaluated on many NLP tasks in three scenarios: few-shot, one-shot and zero-shot. One-shot regards a single demonstration of the task, while zero-shot allows only for natural language instructions, forcing the model to rely solely on the pretraining. In the less constrained few-shots scenario, the GPT-3 was able to surpass the state-of-the-art, composed mostly of fine-tuned models, in some of the tasks. Remarkably, on the NLG task, GPT-3 produced news articles (up to 500 words), which were hard to distinguish from news written by humans.

Domain knowledge. To date, NLP is mostly centered around algorithms that can be trained on available task-specific labeled and unlabeled training samples.¹ In contrast, humans rely on past structured knowledge of the world when facing new challenges. Recently, the search for models that encode the meaning of text, in tasks such as reading comprehension, led the NLP community toward task agnostic models. The popularity of multitask benchmarks (GLUE) is a consequence of this change in focus. Properly encoding domain knowledge is important toward creating task agnostic models that actually map meaning, instead of an “empty” statistical prediction.⁷ We identified two promising

trends to provide domain knowledge on top of the training procedure: KGs and Knowledge Distillation (KD).

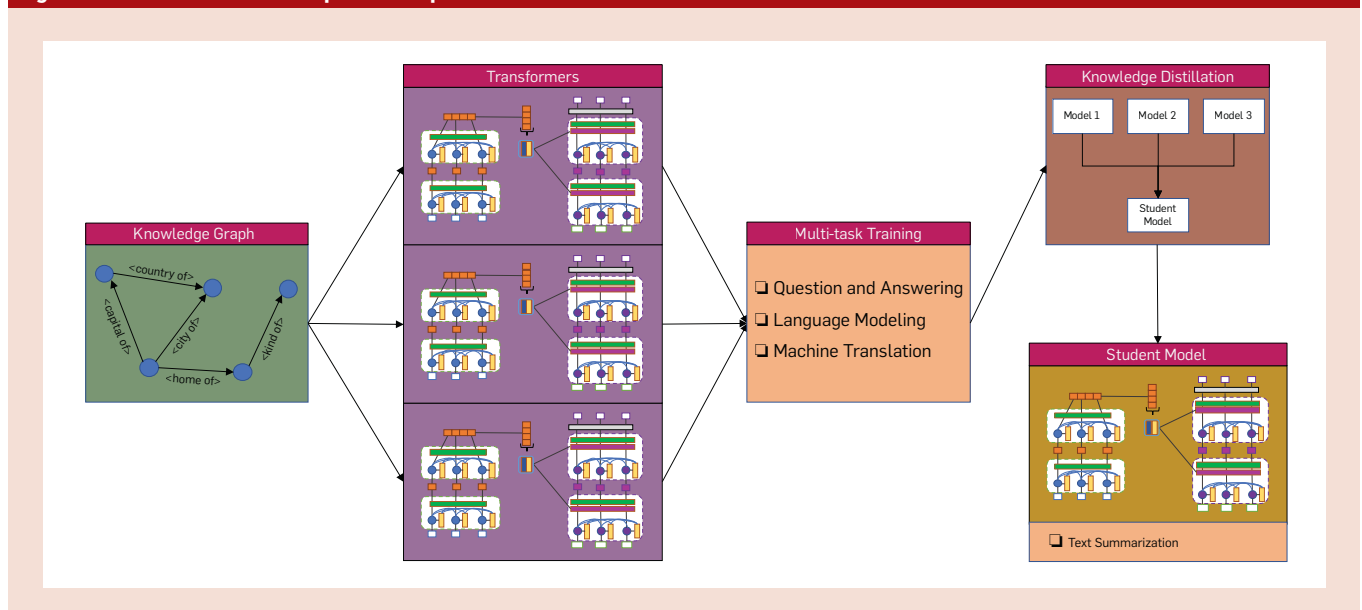
Knowledge graphs. To provide prior facts for DNNs, the community recurs to KGs. Within the KGs framework, knowledge means an organized set of world information, and prior facts are mapped as a graph of object state changes. As a clear advantage, when a DNN has access to prior knowledge it can be trained with less labeled data. For instance, triplets of $\langle object, relation, object \rangle$ as in $\langle Italy, capital, Rome \rangle$, where the objects are the vertices of a KG and the relations are its edges.¹ Moreover, by leveraging prior world knowledge the model can avoid storing these straightforward correlations, and spend trainable parameters on complex statistical reasoning.

Combining attention and KGs, An-nervaz et al.¹ proposed the extraction of knowledge through self-attention, in order to reduce the attention search space. They trained a multitask model following the supervised learning paradigm, using $\max p(y|x, xw, \theta)$ as objective function, where θ is the set of model weights, and the optimization process also factor in the world knowledge input xw . It outperformed similar architectures trained fully on labeled data. Das et al.⁷ evince another key advantage of KGs: enabling the use of the graph theory toolbox. They propose a model based on KGs for reading comprehension. KG is framed as a dynamic

memory model that form a graph correlating objects to its location in the world and clustering them by spatial proximity. For each input sentence, the model queries the state of all objects and propagate changes to all other nodes. After a series of ablation studies, the authors show that their KG encodes enough knowledge priors to achieve state-of-the-art performance. Noteworthy, the model learned commonsense constraints from data, without manual interference. An alternative way to impose priors to the training was proposed by Shirish Keskar et al.,²⁸ in the CTRL model. Their target application is NLG. It is a 1.63B parameters pre-trained Transformer, conditioned to control codes. These codes refers to actual labels defining additional features of text segments. For instance, text style, topic, date or entities’ names.

Knowledge distillation. New neural architectures are constantly surpassing previous ones and the complexity of such models increases as fast. In contrast, these networks are impractical on mobile or real-time scenarios, which represent many real-world applications with limited resources.³² KD is the process of compressing the knowledge of a huge model (teacher) into a lighter representation (student), while minimizing the performance loss. Differently from KG, the KD transfers knowledge by teaching the student model to mimic the teacher and yield a similar prediction. In a general form,

Figure 4. Architecture for a complete and up-to-date NLP model.



the student model optimizes the negative log-likelihood, given a set of trainable parameters θ_t , and the set of teacher parameters θ_t , the loss function use:¹⁹

$$L_s(\theta_s) = -\sum_x Q(y|x, \theta_t) \log p(y|x, \theta_s) \quad (5)$$

where (x,y) is the set of training instances (inputs and ground truth labels) and $Q(y|x, \theta_t)$ is the distribution yield by the teacher model over the whole training set. When the KD is based on an ensemble, the distribution is given by the average of all n teacher models.

Following the framework proposed by You et al.³⁹ for training BERT models, the DistilBERT²⁷ applies knowledge distillation to the pre-training step. The key idea is to yield a pre-trained BERT model with reduced size, which is achieved by keeping 97% of the model performance on downstream tasks, at 40% (student model) of its size. Their method proposed a loss function composed of three potentials: masked language modeling (original BERT), distillation loss, and a cosine distance to align context vectors between the student and teacher models.

Shallow models. Tang et al.³² argues that shallow neural models are not obsolete and can achieve results on par with their very deep counterparts through KD. They experiment using the original BERT model fine-tuned for machine translation as a teacher, and slightly augment the dataset with

synthetic data. The key difference is the use of a single-layer bidirectional LSTM as student model, which not only compresses way less parameters, but is also architecturally different. These discrepancies assert that KD is indeed a model-agnostic procedure, having direct impact only on the objective function. More remarkably, the shallow bidirectional LSTM achieved comparable results to ELMo, while having 100 times fewer parameters and 15-fold faster inference.

Ensembles. With respect to ensembles, in the work of Liu et al.,¹⁸ an ensemble of MT-DNNs reached state-of-the-art performance on GLUE tasks. Although ensembles provide improved generalization and good benchmark scores, they are inadequate for some applications due to the huge size of the complete model. For example, an ensemble of GPT-2 models would require an unreasonable number of parameters for current standards. Therefore, Liu et al.¹⁷ applied KD to generate a single model capable of maintaining the ensemble scores on 7 out of 9 GLUE tasks, outperforming previous single models by a large margin. The authors apply the technique proposed by Hinton et al.¹² to distill the knowledge from the MT-DNNs ensemble. Initially, they trained one ensemble model for each task, and created soft targets for each of the training instances. Soft targets are the average of the predictions of all the individual DNNs on the ensemble.

Next, soft targets are used to inform the student model about how the ensemble generalizes.¹²

Discussion

We compared current state-of-the-art models by GLUE score and number of parameters on the accompanying table. Enabling the training of models such as the GPT-3, two orders of magnitude larger than the already massive BERT, is the latest milestone achieved in recent literature. The table shows large models at the top of current benchmarks, ALBERT being the exception. Novel benchmarks, improving upon GLUE, have been released for future research: SuperGLUE³⁵ proposes harder language understanding tasks, and XGLUE¹⁶ enable cross-lingual model evaluation and training, by providing labels for every task on multiple languages. Lan et al.¹⁵ suggests as an alternative venue exploring additional representation power, that is, engineering the self-supervised losses and proxy tasks in order to map additional dimensions of the data. Given the score achieved by ALBERT, parameter sharing methods can greatly improve BERT-based models efficiency per-parameter, while also imposing additional regularization. Moreover, Tang et al.³² argue that shallow models, such as LSTMs, are capable of more expressiveness than they currently yield, if conditioned to more robust training procedures and KD. Noteworthy, most models listed on the table are based on the original BERT architecture. Yet, the XLnet combined advantages of two successful approaches, BERT and GPT-2, pointing out that AR and AE models should be explored equally in hybrid solutions.

A clear open issue regards commonsense reasoning. As noticed by the Brown et al.,³ increasing model size shows diminishing returns on commonsense reasoning tasks. Das et al.⁷ argues that KGs are a valid solution for augmenting models with commonsense. There is a clear research opportunity on augmenting the proposed KG with other types of object relations, besides world location. With respect to data augmentation, adding finer-grained metadata (for example, control codes²⁸) for each sequence imposes commonsense reasoning on current

Comparing Transformer-based models by score on the GLUE benchmark. Higher reported scores among listed papers, or GLUE's leaderboard. Number of parameters from Sanh et al.²⁷ when not available in the original paper. Score* refers to averaged scores over a subset of the GLUE tasks.

Method	Year	GLUE Score	Params.
Peters et al. ²³	2018	68.7	94M
Sanh et al. ²⁷	2019	77.0	66M
Devlin et al. ⁹	2018	79.5	340M
Liu et al. ¹⁸ (ensemble)	2019	87.6	340M/model
Yang et al. ³⁸	2019	88.4	340M
Liu et al. ²⁰	2019	88.5	355M
Liu et al. ¹⁷	2019	89.9	340M
Lan et al. ¹⁵	2020	90.6	235M
Shoeybi et al. ²⁹	2019	92.0*	3.9B
Radford et al. ²⁵	2019	—	1.5B
Shirish Keskar et al. ²⁸	2019	—	1.63B
Turing-NLG	2020	—	17B
Brown et al. ³	2020	—	175B

models. However, adding prior knowledge to the data could also be harmful. Imposing biases to the raw data, assuming a pretrained model, and then proceeding through multiple training routines for fine-tuning, may yield unexpected outcomes in regard to ethical concerns.²⁶ Therefore, the degree of interpretability associated with novel architectures is a valuable criteria, as happened to the Transformer's self-attention.

Sample efficiency is another concern, since pretraining requires more text than a human would have access to in a lifetime.³ In turn, the performance on most few-shot scenarios is way lower than human baselines. It further highlights the need for additional knowledge encoding methods and remains as an open issue. Consequently, few-shot learning is a huge research trend and will be the main focus of the community for the years to come. On the other hand, as argued by Brown et al.,³ zero- or one-shot scenarios are the closer benchmarks to actual humans.

Conclusion

This article has no intention to lessen the groundbreaking contributions of current literature to the NLP research. From attention to unsupervised training, work as the original Transformers, BERT and GPT-2 were a huge step toward NLU. Instead, we highlight open issues and propose challenges that must be faced henceforward. Attention, mainly self-attention, is indeed a standard on current literature, being part of every recent model reviewed. Nonetheless, to achieve NLU with meaningful models, attention is not enough. Future research efforts can either explore multitask ensemble knowledge in a lighter format or use structured knowledge priors. Extracting knowledge from ensembles rely mostly on KD, which can also be used in a self-learning paradigm, having the ensemble being both the teacher and the student. On the other hand, structured knowledge could also be encoded in a plethora of ways, yet, we argue that KGs still have much uncovered potential. As a final insight, we highlight the potential of older methods, such as LSTMs and GRU, growing on par with DNNs due to multitask learning, knowledge transfer and knowledge

priors. We conjecture that DNNs themselves can still benefit further from these techniques. □

References

- Annervaz, K.M., Chowdhury, S.B.R. and Dukkupati, A. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. *NAACL-HLT*, 2018.
- Bahdanau, D., Cho, K. and Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR abs/1409.0473*, 2014.
- Brown, T.B.B. et al. Language models are few-shot learners. 2020; *arXiv:2005.14165* (2020).
- Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of the 2014 Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics, Doha, Qatar, 103–111; <https://doi.org/10.3115/v1/W14-4012>
- Cho, K. et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation; *arXiv:1406.1078* (2014).
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., and Salakhutdinov, R. Transformer-XL: Attentive language models beyond a fixed-length context. *ACL* (2019).
- Das, R., Munkhdalai, T., Yuan, X., Trischler, A. and McCallum, A. Building dynamic knowledge graphs from text using machine reading comprehension; *arXiv:1810.05682* (2018).
- Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J. and Kaiser, L. Universal transformers; *arXiv:1807.03819* (2018).
- Devlin, J., Chang, M-W, Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conf. North American Chapter of the ACL: Human Language Technologies 1*. Association for Computational Linguistics, Minneapolis, MN, 4171–4186; <https://doi.org/10.18653/v1/N19-1423>
- Gehring, J., Auli, M., Grangier, D., Yarats, D. and Dauphin, Y.N. Convolutional sequence to sequence learning. In *Proceedings of the 34th Intern. Conf. Machine Learning 70*. JMLR, org, 2017, 1243–1252.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018); <https://doi.org/10.1145/3236009>
- Hinton, G., Vinyals, O. and Dean, J. Distilling the knowledge in a neural network. In *Proceedings of the 2015 NIPS Deep Learning and Representation Learning Workshop*.
- Jain, S. and Wallace, B.C. Attention is not explanation. *NAACL-HLT*, 2019.
- Kitchenham, B. and Charters, S. *Guidelines for Performing Systematic Literature Reviews in Software Engineering*. Technical Report. Keele University, 2007.
- Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the Intern. Conf. Learning Representations*. (2020)
- Liang, Y. et al. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *To be published*; <https://bit.ly/3m1OLW7>
- Liu, X., He, P., Chen, W. and Gao, J. Improving multi-task deep neural networks via knowledge distillation for natural language understanding; *arXiv:1904.09482* (2019).
- Liu, X., He, P., Chen, W. and Gao, J. Multi-task deep neural networks for natural language understanding. *ACL*, 2019.
- Liu, Y., Che, W., Zhao, H., Qin, B. and Liu, T. Distilling knowledge for search-based structured prediction. In *Proceedings of the 56th Annual Meeting of the ACL 1*. Association for Computational Linguistics, 2018, Melbourne, Australia, 1393–1402; <https://doi.org/10.18653/v1/P18-1129>
- Liu, Y. et al. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR abs/1907.11692* (2019).
- Lu, J., Batra, D., Parikh, D. and Lee, S. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019; *arXiv:cs.CV/1908.02265*
- Mikolov, T., Chen, K., Corrado, G.S. and Dean, J. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781* (2013).
- Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conf. North American Chapter of the ACL: Human Language Technologies 1*. Association for Computational Linguistics, New Orleans, LA, 2227–2237; <https://doi.org/10.18653/v1/N18-1202>
- Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. Improving language understanding by generative pre-training, 2018; <https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper> (2018).
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskeve, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019).
- Rajani, N.F., McCann, B., Xiong, C. and Socher, R. Explain yourself! Leveraging language models for commonsense reasoning. *ACL*, 2019.
- Sanh, V., Debut, L., Chaumond, J. and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter; *arXiv:1910.01108* (2019).
- Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C. and Socher, R. CTRL: A conditional transformer language model for controllable generation, 2019, *arXiv:1909.05858*.
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J. and Catanzaro, B. Megatron-LM: Training multi-billion parameter language models using model parallelism, 2019; *arXiv:cs.CL/1909.08053*
- Song, K., Tan, X., Qin, T., Lu, J. and Liu, T.-Y. MASS: Masked sequence to sequence pre-training for language GGeneration. *ICML*, 2019; <https://bit.ly/3j90xMN>
- Sutskever, I., Vinyals, O. and Le, Q.V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 2014, 3104–3112.
- Tang, T., Lu, Y., Liu, L., Mou, L., Vehtomova, O. and Lin, J. Distilling task-specific knowledge from BERT into simple neural networks. *arXiv:1903.12136* (2019).
- Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 5998–6008.
- Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *ACL*, 2019.
- Wang, A. et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *CoRR abs/1905.00537* (2019). *arXiv:1905.00537* <http://arxiv.org/abs/1905.00537>
- Wang, A., Singh, A., Michael, J., Hill, F. Levy, O. and Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR abs/1804.07461* (2018).
- Wu, Y. et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR abs/1609.08144* (2016).
- Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R. and Le, Q.V. XLNet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 2019.
- You, Y. et al. Large batch optimization for deep learning: Training BERT in 76 minutes. In *Proceedings of the 2019 Intern. Conf. Learning Representations*.
- Yu, A.W. et al. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv:1804.09541* (2018).

Eduardo Souza Dos Reis is a researcher at SoftwareLab, Unisinos, Brazil.

Cristiano André Da Costa is a professor at SoftwareLab, Unisinos, Brazil.

Diógenes Eugênio Da Silveira is a researcher at SoftwareLab, Unisinos, Brazil.

Rodrigo Simon Bavaresco is a researcher at SoftwareLab, Unisinos, Brazil.

Rodrigo Da Rosa Rigbi is an assistant professor at SoftwareLab, Unisinos, Brazil.

Jorge Luis Victória Barbosa is a professor at SoftwareLab, Unisinos, Brazil.

Rodolfo Stoffel Antunes is an assistant professor at SoftwareLab, Unisinos, Brazil.

Márcio Miguel Gomes is a researcher at SoftwareLab, Unisinos, Brazil.

Gustavo Federizzi is senior manager at Dell Inc., Brazil.

Copyright held by authors/owners.

research highlights

P. 165

**Technical
Perspective**

The Strength of SuRF

By Stratos Idreos

P. 166

Succinct Range Filters

By Huanchen Zhang, Hyeontaek Lim, Viktor Leis, David G. Andersen,
Michael Kaminsky, Kimberly Keeton, and Andrew Pavlo

Technical Perspective

The Strength of SuRF

By Stratos Idreos

DATA STRUCTURES THAT filter data for point or range queries are prevalent across all data-driven applications, from analytics to transactions, and modern machine learning applications. The primary objective is simple: find whether one or more data items exist in the database. Yet, this simple task is exceptionally difficult to perform efficiently, and surprisingly critical for the overall properties of the data-intensive applications that rely on filtering.


This is a hard problem as there are numerous critical parameters and trade-offs. Many parameters come from the workload, for example, the exact percentage of point queries versus updates, percentage of empty-result queries, and so on. Other parameters come from the underlying hardware; for example, filters typically reside in memory but, with exponentially increasing data sizes, we need to be mindful of the filter size and the memory hierarchy. Overall, there are complex trade-offs to navigate: memory, read, and write amplification. For example, a data structure cannot be efficient for both point and range queries while also supporting efficient writes. Yet, numerous applications need to expose both read patterns.

A prototypical application of filters is storage engines based on Log-Structured Merge Trees (LSM-tree). An LSM-tree stores data in the order they arrive in immutable files and periodically sort-merges them into larger files. This way, it behaves in between a log and a sorted array, providing a good balance of read and write performance depending on the exact tuning (file size, buffer size, among others). LSM-tree storage engines are used as the backbone of most distributed key-value stores and applications range from social media, Web-applications, e-shopping, IoT, and so on. Due to their multilevel architecture enforcing a global temporal order, LSM-tree engines rely heavily on in-memory filters.

The trie-based design (of SuRF) enables the construction of a filter that can support performant range queries, point queries, and approximate counts.

Succinct Range Filter (SuRF) was introduced at the ACM SIGMOD 2018 Conference as a new succinct filter.¹ SuRF is based on a trie-like structure termed “Fast Succinct Trie.” The trie-based design enables the construction of a filter that can support performant range queries, point queries, and approximate counts. The authors of SuRF make the following critical and insightful observation that brings everything together and allows SuRF to balance the various hardware and workload trade-offs. For a given set of queries, the upper levels of the trie incur many more accesses than the lower levels. For this reason, the SuRF design utilizes a dense, performance-optimized encoding scheme for the top of the trie and a sparse, memory-optimized encoding scheme for the bottom. This results in a data structure that is both fast and memory efficient. The upper levels, which are comprised of few nodes but incur many accesses, encode keys under a new scheme called LOUDS-Dense which sacrifices space efficiency for fast lookups. The lower levels, which contain the majority of nodes but have a sparser access

pattern are encoded with a scheme called LOUDS-Sparse, which sacrifices fast lookups for space efficiency.

Compared to state-of-the-art bloom filter-based solutions (for example, prefix bloom filters) SuRF provides a general solution, that is, it can support any range query as well as efficient point queries. Compared to state-of-the-art tree or trie-based solutions the trie-based design enables the construction of a filter that can support performant range queries, point queries, and approximate counts. SuRF offers similar or better performance at a much smaller memory footprint. The SuRF paper that follows shows end-to-end impact by integrating SuRF in RocksDB, the most mature LSM-tree based storage engine, and demonstrating strong results (for example, up to 5x) in time-series applications for both point and range queries. SuRF can be applied broadly to any application that needs a succinct filter such as monitoring, privacy/security, and graph analytics. Finally, the core spirit of the design of SuRF exemplifies elegant research taste in pursuing hybrid, hardware- and workload-conscious designs. 

Reference

1. Zhang, H., Lim, H., Leis, V., Andersen, D.G., Kaminsky, M., Keeton, K. and Pavlo, A. SuRF: Practical range query filtering with fast succinct tries. In *Proceedings of ACM SIGMOD 2018*.

Stratos Idreos is an associate professor of computer science at the Harvard John A. Paulson School of Engineering and Applied Science. He leads the Data Systems Laboratory at Harvard SEAS, Cambridge, MA, USA.

Copyright held by author.

Succinct Range Filters

By Huanchen Zhang, Hyeontaek Lim, Viktor Leis, David G. Andersen, Michael Kaminsky, Kimberly Keeton, and Andrew Pavlo

Abstract

We present the *Succinct Range Filter* (SuRF), a fast and compact data structure for approximate membership tests. Unlike traditional Bloom filters, SuRF supports both single-key lookups and common range queries, such as range counts. SuRF is based on a new data structure called the *Fast Succinct Trie* (FST) that matches the performance of state-of-the-art order-preserving indexes, while consuming only 10 bits per trie node—a space close to the minimum required by information theory. Our experiments show that SuRF speeds up range queries in a widely used database storage engine by up to 5×.

1. INTRODUCTION

Write-optimized log-structured merge (LSM) trees¹⁶ are popular low-level storage engines for general-purpose databases that provide fast writes^{1, 14, 18} and ingest-abundant DBMSs such as time-series databases.^{5, 17} One of their main challenges for fast query processing is that items could reside in different immutable files (SSTables) from all levels. Item retrieval in these systems may therefore incur multiple expensive disk I/Os.^{16, 18}

Many LSM tree-based systems use Bloom filters to “guard” on-disk files to reduce the number of unnecessary I/Os^{2, 3, 17, 18}: they read an on-disk file only when the associated in-memory Bloom filter indicates that the query item may exist in the file. Bloom filters are a good match for this task. First, Bloom filters are fast and small enough to reside in memory. Second, Bloom filters answer approximate membership tests with “one-sided” errors—if the querying item is a member, the filter is guaranteed to return true; otherwise, the filter will likely return false, but may incur a false positive.

Although Bloom filters are useful for single-key lookups (“Is key 50 in the SSTable?”), they cannot handle range queries (“Are there keys between 40 and 60 in the SSTable?”). With only Bloom filters, an LSM tree-based storage engine still needs to read additional disk blocks for range queries. Alternatively, one could maintain an auxiliary index, such as a B+Tree, to accelerate range queries, but the memory cost would be significant. To partly address the high I/O cost of range queries, LSM tree-based designs often use *prefix Bloom filters* to optimize certain fixed-prefix queries (e.g., “where email starts with com.foo@”),^{2, 11, 17} despite their inflexibility for more general range queries.

To address these limitations, we present the **Succinct Range Filter** (SuRF), a fast and compact data structure that provides exact-match filtering, range filtering, and approximate range counts. Like Bloom filters, SuRF guarantees one-sided errors for point and range membership tests. SuRF can trade between false positive rate and memory consumption, and this trade-off is tunable for point and

range queries semi-independently.

SuRF is built upon a new space-efficient data structure called the *Fast Succinct Trie* (FST). It performs comparably to or better than state-of-the-art uncompressed index structures for both integer and string workloads. FST consumes only 10 bits per trie node, which is close to the information-theoretic lower bound.

The key insight in SuRF is to transform the FST into an approximate (range) membership filter by removing levels of the trie and replacing them with some number of suffix bits. The number of such bits (either from the key itself or from a hash of the key—as we discuss later in the paper) trades space for decreased false positives.

We evaluate SuRF via microbenchmarks and as a Bloom filter replacement in RocksDB—a widely-used database storage engine.² Our experiments on a 100GB time-series dataset show that replacing the Bloom filters with SuRFs of the same filter size reduces I/O. This speeds up closed-range queries (i.e., with an upper bound) by up to 5× compared to the original implementation, with a modest cost on the worst-case point query throughput due to slightly higher false positive rate. One can eliminate this performance gap by increasing the size of SuRFs by a few bits per key.

2. FAST SUCCINCT TRIES

The core data structure in SuRF is the *Fast Succinct Trie* (FST). FST is a space-efficient, static trie that answers point and range queries. FST is 4–15× faster than earlier succinct tries,^{6, 12} achieving performance comparable to or better than the state-of-the-art pointer-based indexes.^{8, 15, 19}

FST’s design is based on the observation that the upper levels of a trie comprise few nodes but incur many accesses. The lower levels comprise the majority of nodes, but are relatively “colder.” We therefore encode the lower levels of the trie using the succinct **LOUDS-Sparse** scheme to guarantee the overall space-efficiency of the data structure. Here, LOUDS stands for the *Level-Ordered Unary Degree Sequence*.¹³ By contrast, we encode the upper levels using a fast bitmap-based encoding scheme, called **LOUDS-Dense**, in which a child node search requires only one array lookup, choosing performance over space.

For the rest of the section, we assume that the trie maps the keys to fixed-length values. We also assume that the trie has a fanout of 256 (i.e., one byte per level).

The original version of this paper is entitled “SuRF: Practical Range Query Filtering with Fast Succinct Tries” and was published in *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data* (Houston, TX, USA).

2.1 LOUDS-Sparse

LOUDS-Sparse encodes the trie nodes in the level order using four byte-/bit-sequences, as shown in the lower half of Figure 1.

The first byte-sequence, *S-Labels*, records all the branching labels for each trie node. As an example, the second node at level 2 in Figure 1 has three branches. *S-Labels* includes its labels *r*, *s*, and *t* in order. We denote the case where the prefix leading to a node is also a valid key using the special byte $0 \times \text{FF}$ ¹ at the beginning of the node. For example, in Figure 1, the second node at level 3 has ‘fas’ as its incoming prefix. As ‘fas’ itself is also a key stored in the trie, the node adds $0 \times \text{FF}$ to *S-Labels* as the first byte. Because the special byte always appears at the beginning of a node, it can be distinguished from the real $0 \times \text{FF}$ label.

The second bit-sequence (*S-HasChild*) includes one bit for each byte in *S-Labels* to indicate whether a child branch continues (i.e., points to a subtree) or terminates (i.e., points to a value). Taking the rightmost node at level 2 in Figure 1 as an example, because the branch labeled *i* points to a subtree, the corresponding bit in *S-HasChild* is set. The branch labeled *y*, however, points to a value, and its *S-HasChild* bit is cleared.

The third bit-sequence (*S-LOUDS*) denotes node boundaries: if a label is the first in a node, its *S-LOUDS* bit is set. Otherwise, the bit is cleared. For example, in Figure 1, the second node at level 2 has three branches and is encoded as 100 in *S-LOUDS*.

The final byte-sequence (*D-Values*) stores the fixed-length values (e.g., pointers) mapped by the keys. The values are concatenated in the level order—same as the three bitmaps.

Tree navigation relies on the fast rank & select primitives. Given a bit-vector, $\text{rank}(i)$ counts the number of 1s up to position i , while $\text{select}(i)$ returns the position of the i th 1. Modern rank & select implementations such as as^{21} achieve *constant time* by using lookup tables to store a sampling of precomputed results so that they only need to count between the samples. We denote rank/select over bit-sequence bs on position pos to be $\text{rank/select}(bs, pos)$.

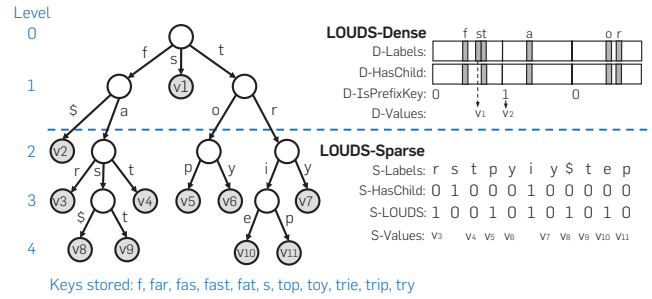
Let pos be the current bit position in *S-Labels*. Assume that $\text{S-HasChild}[pos] = 1$, indicating that the branch at pos points to a child node. To move to the child node, we first compute the child node’s rank in the overall level-ordered node list: $r = \text{rank}(\text{S-HasChild}, pos) + 1$. Because every node only has its first bit set in *S-LOUDS*, we can use $\text{select}(\text{S-LOUDS}, r)$ to find the position of that child node.

To move to the parent node, we first get the rank r of the current node by $r = \text{rank}(\text{S-LOUDS}, pos)$ because the number of ones in *S-LOUDS* indicates the number of nodes. We then find the node that contains the $(r - 1)$ th children: $\text{select}(\text{S-HasChild}, r - 1)$.

Given $\text{S-HasChild}[pos] = 0$, to access the associated value, we compute its index in *S-Values*. Because every cleared bit in *S-HasChild* has a value, there are $pos - \text{rank}(\text{S-HasChild}, pos)$ values before pos .

¹ If a node has a single branching label $0 \times \text{FF}$, it must be the real $0 \times \text{FF}$ byte (otherwise, the node will not exist in the trie).

Figure 1. An example fast succinct trie. The upper and lower levels of the trie are encoded using LOUDS-Dense and LOUDS-Sparse, respectively. “\$” represents the character whose ASCII number is $0 \times \text{FF}$. It is used to indicate the situation where the prefix leading to a node is also a valid key.



2.2 LOUDS-Dense

As shown in the top half of Figure 1, LOUDS-Dense encodes each trie node using three bitmaps of size 256 and a byte-sequence to hold the values. The encoding follows the level order.

The first bitmap (*D-Labels*) records the branching labels for each node. Specifically, the i th bit in the bitmap ($0 \leq i \leq 255$) indicates whether the node has a branch with label i . For example, the root node in Figure 1 has three outgoing branches labeled *f*, *s*, and *t*. The *D-Labels* bitmap thus sets the 102nd (*f*), 115th (*s*), and 116th (*t*) bits and clears the rest.

The second bitmap (*D-HasChild*) indicates whether a branch points to a subtree or terminates (i.e., points to the value or the branch does not exist). Taking the root node in Figure 1 as an example, the *f* and the *t* branches continue with subtrees, while the *s* branch terminates with a value. In this case, the *D-HasChild* bitmap only sets the 102nd (*f*) and 116th (*t*) bits for the node.

The third bitmap (*D-IsPrefixKey*) includes only one bit per node to indicate whether the prefix that leads to the node is also a valid key. The same case is handled by the special byte $0 \times \text{FF}$ in LOUDS-Sparse.

The final byte-sequence (*D-Values*) is organized the same way as *S-Values* in LOUDS-Sparse.

Tree navigation in LOUDS-Dense also uses the rank & select primitives. Given a position pos in *D-Labels*, to move to the child node: $256 \times \text{rank}(\text{D-HasChild}, pos)$; to move to the parent: $\text{select}(\text{D-HasChild}, \lfloor pos/256 \rfloor)$; to access the value: $\text{rank}(\text{D-Labels}, pos) - \text{rank}(\text{D-HasChild}, pos) + \text{rank}(\text{D-IsPrefixKey}, \lfloor pos/256 \rfloor) - 1$.

LOUDS-Dense is faster than LOUDS-Sparse because (1) label search within a node requires only one lookup in the bitmap rather than a binary search and (2) move to child only computes a rank in one bit-vector instead of a rank and a select on different bit-vectors.

2.3 FST and operations

FST is a hybrid trie in which the upper levels are encoded with LOUDS-Dense and the lower levels with LOUDS-Sparse. The dividing point between the upper and lower levels is tunable to trade performance and space. By default, we

keep the size ratio R between LOUDS-Dense and LOUDS-Sparse to be less than 1:64 in favor of the space-efficiency provided by LOUDS-Sparse.

FST supports four basic operations efficiently:

- **ExactKeySearch**(key): Return the value of key if key exists (or NULL otherwise).
- **LowerBound**(key): Return an iterator pointing to the key-value pair (k, v) where k is the smallest in lexicographical order satisfying $k \geq key$.
- **MoveToNext**($iter$): Move the iterator to the next key.
- **Count**($lowKey, highKey$): Return the number of keys contained in the range ($lowKey, highKey$).

A point query (i.e., *ExactKeySearch*) in FST works by first searching the LOUDS-Dense levels. If the search does not terminate, it continues into the LOUDS-Sparse levels. The high-level searching steps at each level are similar regardless of the encoding mechanism: First, search the current node’s label sequence for the target key byte. If the key byte does not exist, terminate and return NULL. Otherwise, check the corresponding bit in the *HasChild* bit-sequence. If the bit is set, compute the child node’s starting position in the label sequence and continue to the next level. Otherwise, return the corresponding value in the value sequence.

LowerBound uses a high-level algorithm similar to the point query implementation. Instead of an exact match, the algorithm searches the current node’s label sequence for the smallest label that is greater than or equal to the search byte of that level. The algorithm may recursively move up to the parent node if the search hits node boundaries. Once such label L is found, the algorithm moves iterator to the left-most key in the subtree rooted at L .

We include per-level cursors in the iterator to record a trace from root to leaf (i.e., the per-level positions in the label sequence) for the current key. Using the cursors, range scans (*MoveToNext*) in FST are implemented efficiently. Each level cursor is initialized once through a “move-to-child” call from its upper-level cursor. After that, scan operations at this level only involve cursor movement, which is cache-friendly and fast. Our evaluation shows that range queries in FST are even faster than pointer-based tries.

For *Count*, the algorithm first performs *MoveToNext* on both boundaries and obtains two iterators. It extends each iterator down the trie and sets the cursor at each level to the position of the smallest leaf key that is greater than the current key, until the two iterators meet or reach the maximum trie height. The algorithm then counts the number of leaf nodes at each level between the two iterators by computing the difference of their ranks on the *D-HasChild/S-HasChild* bit-vector. The sum of those counts is returned.

Finally, FST can be built using a single scan over a sorted key-value list.

2.4 Space analysis

A tree representation is “succinct” if the space taken by the representation is close to the information-theoretic lower

bound, which is the minimum number of bits needed to distinguish any object in a set. The information-theoretic lower bound of a trie of degree k is approximately $n(k \log_2 k - (k - 1) \log_2 (k - 1))$ bits (9.44 bits when $k = 256$ in our case).⁷

Given an n -node trie, LOUDS-Sparse uses $8n$ bits for *S-Labels*, n bits for *S-HasChild*, and n bits for *S-LOUDS*, a total of $10n$ bits (plus auxiliary bits for rank & select). Although the space taken by LOUDS-Sparse is close to the information-theoretic lower bound, technically, LOUDS-Sparse can only be categorized as *compact* rather than *succinct* in a finer classification scheme because LOUDS-Sparse takes $O(Z)$ space (despite the small multiplier) instead of $Z + o(Z)$.

LOUDS-Dense’s size is restricted by the size ratio R to ensure that it does not affect the overall space efficiency of FST. Notably, LOUDS-Dense does not always take more space than LOUDS-Sparse: if a node’s fanout is larger than 51, it takes fewer bits to encode the node using the former instead of the latter. As such nodes are common in a trie’s upper levels, adding LOUDS-Dense on top of LOUDS-Sparse often improves space efficiency.

3. SUCCINCT RANGE FILTERS

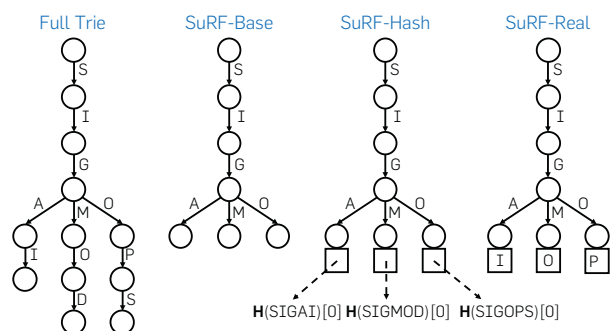
In building SuRF using FST, our goal was to balance a low false positive rate with the memory required by the filter. The key idea is to use a truncated trie, that is, to remove lower levels of the trie and replace them with suffix bits extracted from the key. We introduce three variations of SuRF. We describe their properties and how they guarantee one-sided errors. The current SuRF design is static, requiring a full rebuild to insert new keys.

3.1 Basic SuRF

FST is a trie-based index structure that stores complete keys. As a filter, FST is 100% accurate; the downside, however, is that the full structure can be big. In many applications, filters must fit in memory to guard access to a data structure stored on slower storage. These applications cannot afford the space for complete keys and thus must trade accuracy for space.

The basic version of SuRF (SuRF-Base) stores the minimum-length key prefixes such that it can uniquely identify each key. Specifically, SuRF-Base only stores an additional byte for each key beyond the shared prefixes. Figure 2

Figure 2. SuRF variations. Deriving SuRF variations from a full trie.



shows an example. Instead of storing the full keys ('SIGAI', 'SIGMOD', 'SIGOPS'), SuRF-Base truncates the full trie by including only the shared prefix ('SIG') and one more byte for each key ('C', 'M', 'O').

Pruning the trie in this way affects both filter space and accuracy. Unlike Bloom filters where the keys are hashed, the trie shape of SuRF-Base depends on the distribution of the stored keys. Hence, there is no theoretical upper bound of the size of SuRF-Base. Empirically, however, SuRF-Base uses only 10 bits per key (BPK) for 64-bit random integers and 14 BPK for emails. The intuition is that the trie built by SuRF-Base usually has an average fanout $F > 2$: there are less than twice as many nodes as keys. Because FST (LOUDS-Sparse to be precise) uses 10 bits to encode a trie node, the size of SuRF-Base is less than 20 BPK for $F > 2$.

Filter accuracy is measured by the false positive rate (FPR). A false positive in SuRF-Base occurs when the prefix of the nonexistent query key coincides with a stored key prefix. For example, in Figure 2, querying key 'SIGMETRICS' will cause a false positive in SuRF-Base. FPR in SuRF-Base depends on the distributions of the stored and query keys. Our results in Section 4.2 show that SuRF-Base incurs a 4% FPR for integer keys and a 25% FPR for email keys. To improve FPR, we include two forms of key suffixes described here to allow SuRF to better distinguish between key prefixes.

3.2 SuRF with hashed key suffixes

As shown in Figure 2, SuRF with hashed key suffixes (SuRF-Hash) adds a few hash bits per key to SuRF-Base to reduce its FPR. Let H be the hash function. For each key K , SuRF-Hash stores the n (n is fixed) least-significant bits of $H(K)$ in FST's value array (which is empty in SuRF-Base). When a key (K') lookup reaches a leaf node, SuRF-Hash extracts the n least-significant bits of $H(K')$ and performs an equality check against the stored hash bits associated with the leaf node. Using n hash bits per key guarantees that the point query FPR of SuRF-Hash is less than 2^{-n} (the partial hash collision probability). Experiments in Section 4.2 show that SuRF-Hash requires only 2–4 hash bits to reach 1% FPR.

The extra bits in SuRF-Hash do not help range queries because they do not provide ordering information on keys.

3.3 SuRF with real key suffixes

Instead of hash bits, SuRF with real key suffixes (SuRF-Real) stores the n key bits immediately following the stored prefix of a key. Figure 2 shows an example when $n = 8$. SuRF-Real includes the next character for each key ('I', 'O', 'P') to improve the distinguishability of the keys: for example, querying 'SIGMETRICS' no longer causes a false positive. Unlike in SuRF-Hash, both point and range queries benefit from the real suffix bits to reduce false positives. For point queries, the real suffix bits are used the same way as the hashed suffix bits. For range queries (e.g., move to the next key $> K$), when reaching a leaf node, SuRF-Real compares the stored suffix bits s to key bits k_s of the query key at the corresponding position.

If $k_s \leq s$, the iterator points to the current key; otherwise, it advances to the next key in the trie.

Although SuRF-Real improves FPR for both point and range queries, the trade-off is that using real keys for suffix bits cannot provide as good FPR as using hashed bits because the distribution correlation between the stored keys and the query keys weakens the distinguishability of the real suffix bits.

4. SURF MICROBENCHMARKS

In this section, we evaluate SuRF using in-memory micro-benchmarks to provide a comprehensive understanding of the filter's strengths and weaknesses.

The underlying data structure FST is evaluated separately in our original paper.²⁰ We found that compared to state-of-the-art pointer-based indexes such as B+tree⁸ and the Adaptive Radix Tree (ART),¹⁵ FST matches their performance while being an order-of-magnitude smaller. We also compared FST against other succinct trie alternatives^{6, 12} and showed that FST is 4–15× faster and is also smaller than these previous solutions.

4.1 Experiment setup

We use the YCSB⁹ workloads C and E to generate point and range queries. We test two representative key types: 64-bit random integers generated by YCSB and email addresses (host reversed, e.g., "com.domain@foo") drawn from a real-world dataset (average length = 22 bytes, max length = 129 bytes).

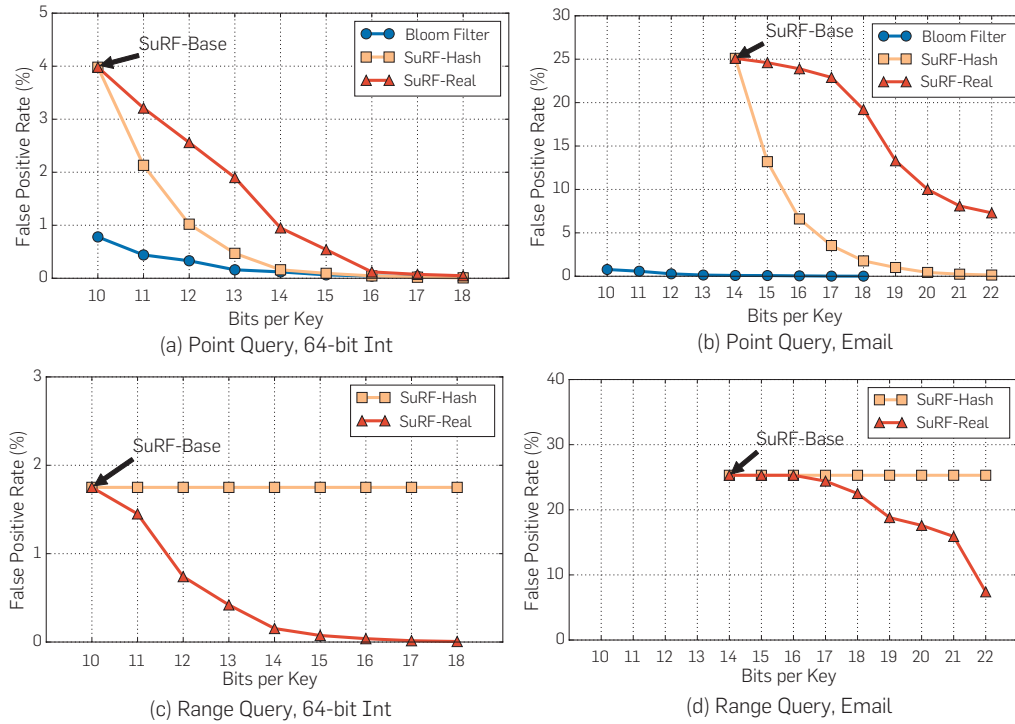
The three most important metrics with which to evaluate SuRF are false positive rate (FPR), performance, and space. The datasets are 100M 64-bit random integer keys and 25M email keys. In the experiments, we first construct the filter under test using half of the dataset selected at random. We then execute 10M point or range queries on the filter. The querying keys (K) are drawn from the *entire* dataset according to YCSB workload C so that roughly 50% of the queries return false. For 64-bit random integer keys, the range query is $[K + 2^{37}, K + 2^{38}]$ where 46% of the queries return true. For email keys, the range query is $[K, K$ (with last byte ++)] (e.g., [org.acm@sigmod, org.acm@sigmoe]) where 52% of the queries return true.

4.2 False positive rate

Figure 3 shows the false positive rate (FPR) comparison between SuRF variants and the Bloom filter by varying the size of the filters. The Bloom filter only appears in point queries. Note that SuRF-Base consumes 14 (instead of 10) bits per key for the email key workloads. This is because email keys share longer prefixes, which increases the number of internal nodes in SuRF.

For point queries, the Bloom filter has lower FPR than the same-sized SuRF variants in most cases, although SuRF-Hash catches up quickly as the number of bits per key increases because every hash bit added cuts the FPR by half. Real suffix bits in SuRF-Real are generally less effective than hash bits for point queries. For range queries, only SuRF-Real benefits from increasing filter size because the hash suffixes in SuRF-Hash do not provide ordering

Figure 3. SuRF false positive rate. False positive rate comparison between SuRF variants and the Bloom filter (lower is better).



information. The shape of the SuRF-Real curves in the email key workloads (i.e., the latter four suffix bits are more effective in recognizing false positives than the earlier four) is because of ASCII encoding of characters.

We also observe that SuRF variants have higher FPRs for the email key workloads. This is because the email keys in the dataset are very similar (i.e., the key distribution is dense). Two email keys often differ by the last byte, or one may be a prefix of the other. If one of the keys is represented in the filter and the other key is not, querying the missing key on SuRF-Base is likely to produce false positives. The high FPR for SuRF-Base is significantly lowered by adding suffix bits, as shown in the figures.

4.3 Performance

Figure 4 shows the throughput comparison. The SuRF variants operate at a speed comparable to the Bloom filter for the 64-bit integer key workloads, thanks to the hybrid encodings and other performance optimizations such as vectorized label search and memory prefetching. For email keys, the SuRF variants are slower than the Bloom filter because of the overhead of searching/traversing the long prefixes in the trie. The Bloom filter’s throughput decreases as the number of bits per key gets larger because larger Bloom filters require more hash probes. The throughput of the SuRF variants does not suffer from increasing the number of suffix bits because as long as the suffix length is less than 64 bits, checking with the suffix bits only involves one memory access and one integer comparison. Range queries in SuRF are slower than point queries because every query needs to walk down to the bottom of the trie (no early exit).

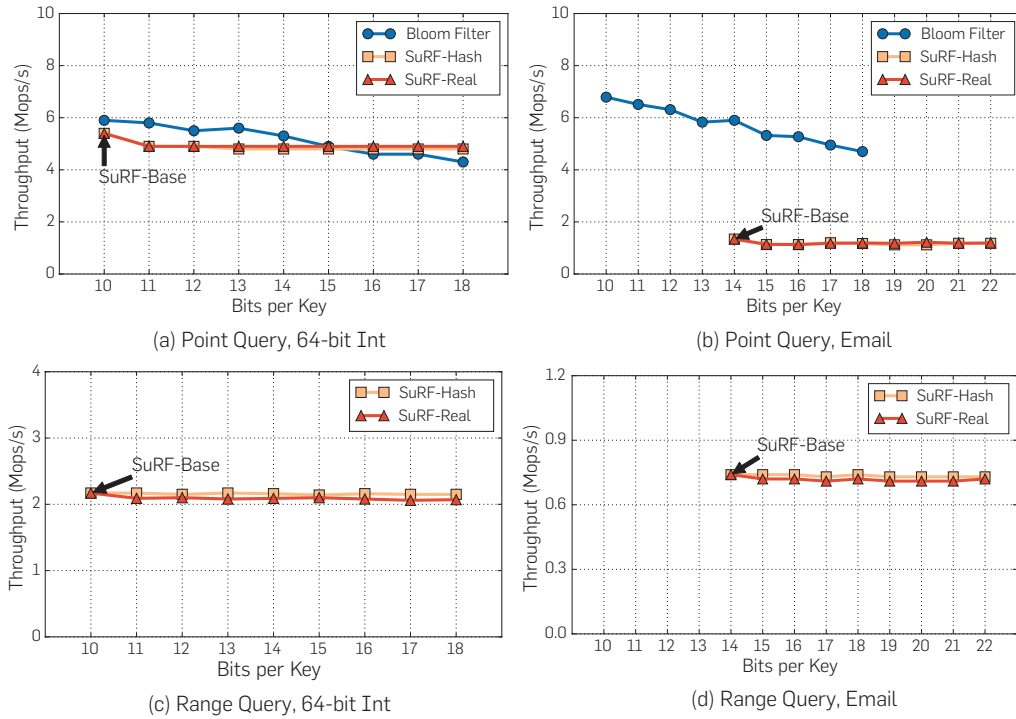
Some high-level takeaways from the experiments are as follows: (1) SuRF can perform range filtering while the Bloom filter cannot. (2) If the target application only needs point query filtering with moderate FPR requirements, the Bloom filter is usually a better choice than SuRF. (3) For point queries, SuRF-Hash can provide similar theoretical guarantees on FPR as the Bloom filter, while the FPR for SuRF-Real depends on the key distribution.

5. EXAMPLE APPLICATION: ROCKSDB

We integrated SuRF with RocksDB as a replacement for its Bloom filter. Incoming writes go into the RocksDB’s MemTable. When the MemTable is full (e.g., exceeds 4MB), the engine sorts it and then converts it into an SSTable at level 0. An SSTable contains sorted key-value pairs and is divided into fixed-length blocks matching the smallest disk access units. To locate blocks, RocksDB stores the “restarting point” (a string that is \geq the last key in the current block and $<$ the first key in the next block) for each block as the block index. When the size of a level hits a threshold, RocksDB selects an SSTable at this level and merges it into the next-level SSTables that have overlapping key ranges. This process is called compaction. The keys are globally sorted across SSTables for each level ≥ 1 . This property ensures that an entry lookup reads at most one SSTable per level for levels ≥ 1 .

We modified RocksDB’s point (*Get*) and range (*Seek*) query implementations to use SuRF. For *Get(key)*, RocksDB uses SuRF exactly like the Bloom filter where at each level, it locates the candidate SSTable(s) and block(s) via the block indexes. For each candidate SSTable, RocksDB queries the

Figure 4. SuRF performance. Performance comparison between SuRF variants and the Bloom filter (higher is better).



in-memory filter first and fetches the SSTable block only if the filter result is positive.

To implement $Seek(lk, hk)$, RocksDB first collects the candidate SSTables from all levels by searching for lk in the block indexes. Absent SuRFs, RocksDB examines each candidate SSTable and fetches the block containing the smallest key that is $\geq lk$. RocksDB then finds the global smallest key $K \geq lk$ among those candidate keys. If $K \leq hk$, the query succeeds; otherwise, the query returns empty.

With SuRFs, however, instead of fetching the actual blocks, RocksDB obtains the candidate key for each SSTable by performing a *LowerBound* query on its SuRF to avoid the one I/O per SSTable. If the query succeeds, RocksDB fetches exactly one block from the selected SSTable that contains the global minimum K . If the query returns empty, no I/O is involved. Because SuRF only stores key prefixes, the system must perform additional checks to break ties and to prevent false positives. The additional checks are described in our original paper.²⁰ Despite those potential checks, using SuRF in RocksDB reduces the average I/Os per $Seek(lk, hk)$ query.

5.1 Evaluation setup

Time-series databases often use RocksDB or similar LSM-tree designs as their storage engine.^{5,17} We thus create a synthetic RocksDB benchmark to model a time-series dataset generated from distributed sensors for our end-to-end performance measurements. We simulated 2k sensors to record events. The key for each event is a 128-bit value comprised of a 64-bit timestamp followed by a 64-bit

sensor ID. The associated value in the record is 1KB long. The occurrence of each event detected by each sensor follows a Poisson distribution with an expected frequency of one every 0.2 s. Each sensor operates for 10K seconds and records $\sim 50K$ events. The starting timestamp for each sensor is randomly generated within the first 0.2 s. The total size of the raw records is approximately 100GB.

Our testing framework supports the following queries:

- **Point Query:** Given a timestamp and a sensor ID, return the record if there is an event.
- **Range Query:** Given a time range, determine whether any events happened during that time period. If yes, return an iterator pointing to the earliest event in the range.

Our test machine has an Intel Core i7-6770HQ CPU, 32 GB RAM, and an Intel 540s 480GB SSD. We configured² RocksDB according to Facebook’s recommendations.^{4,11} The resulting RocksDB instance has four levels and uses 52GB of disk space.

We create four instances of RocksDB with different filter options: no filter, Bloom filter, SuRF-Hash, and SuRF-Real. We configure each filter to use an equal amount of memory. Bloom filters use 14 bits per key. The equivalent-sized SuRF-Hash and SuRF-Real include a 4-bit suffix per key. We first warm the cache with 1 million uniformly-distributed point queries to existing keys so that every SSTable is touched roughly 1000 times, and the block indexes and

² Block cache size = 1 B; OS page cache $\leq 3GB$.

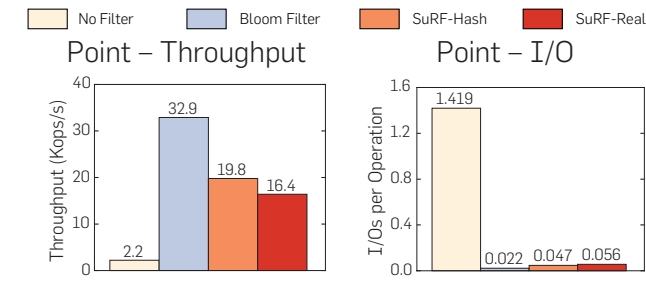
filters are cached. After the warm-up, both RocksDB’s block cache and the OS page cache are full. We then execute 50k application queries, recording the DBMS’s end-to-end throughput and I/O counts. The query keys (for range queries, the starting keys) are randomly generated: a random timestamp within the operated time range + a randomly picked sensor ID. The reported numbers are the average of three runs.

5.2 Point query results

Figure 5 shows the result for point queries. Because the query keys are randomly generated, almost all queries return false. The query performance is dominated by the I/O count: they are inversely proportional. Excluding Level 0, each point query is expected to access three SSTables, one from each level (Level 1, 2, 3). Without filters, point queries incur approximately 1.5 I/Os per operation according to Figure 5, which means that the entire Level 1 and approximately half of Level 2 are likely cached. This is representative of typical RocksDB configurations where the last two levels are not cached in memory.¹⁰

Using filters in point queries reduces I/O because they prevent unnecessary block retrieval. Using SuRF-Hash or SuRF-Real is slower than using the Bloom filter because the 4-bit suffix does not reduce false positives as low as the Bloom filter configuration (refer to Section 4.2). SuRF-Real provides similar benefit to SuRF-Hash because the key distribution is sparse.

Figure 5. Point queries. RocksDB point query evaluation under different filter configurations.



5.3 Range query results

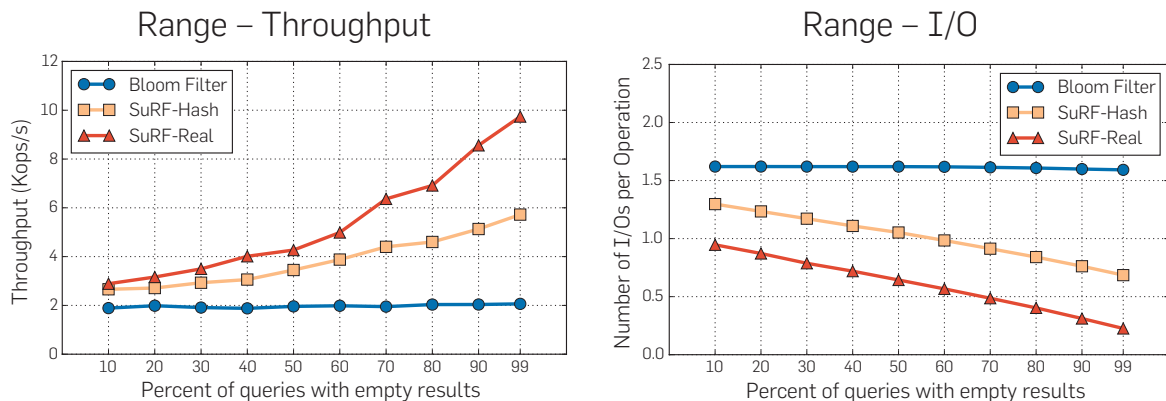
The main benefit of using SuRF is speeding up range queries. Figure 6 shows the throughput and I/O count for range queries. On the x-axis, we control the percentage of queries with empty results by varying the range size. The Poisson distribution of events from all sensors has an expected frequency of one per $\lambda = 10^5$ ns. For an interval with length R , the probability that the range contains no event is given by $e^{-R/\lambda}$. Therefore, for a target percentage (P) of Closed-Seek queries with empty results, we set range size to $\lambda \ln\left(\frac{1}{P}\right)$. For example, for 50%, the range size is 69310 ns.

As shown in Figure 6, the Bloom filter does not help range queries and is equivalent to having no filter. Using SuRF-Real, however, speeds up the query by 5× when 99% of the queries return empty. Again, I/O count dominates performance. Without a range filter, every query must fetch candidate SSTable blocks from each level to determine whether there are keys in the range. Using the SuRF variants, however, avoids many of the unnecessary I/Os; RocksDB performs a read to the SSTable block containing that minimum key only when the minimum key returned by the filters at each level falls into the querying range. Using SuRF-Real is more effective than SuRF-Hash because the real suffix bits help reduce false positives at the range boundaries.

6. CONCLUSION

This paper introduces the SuRF filter structure, which supports approximate membership tests for single keys and ranges. SuRF is built upon a new succinct data structure, called the Fast Succinct Trie (FST), that requires only 10 bits per node to encode the trie. FST is engineered to have performance equivalent to state-of-the-art pointer-based indexes. SuRF is memory efficient, and its space and false positive rates can be tuned by choosing different amounts of suffix bits to include. Replacing the Bloom filters with SuRFs of the same size in RocksDB substantially reduced I/O and improved throughput for range queries with a modest cost on the worst-case point query throughput. We believe, therefore, that SuRF is a promising technique for optimizing future storage systems, and more. SuRF’s source code is publicly available at <https://github.com/efficient/SuRF>. □

Figure 6. Range queries. RocksDB range query evaluation under different filter configurations and range sizes.



References

1. Facebook MyRocks. <http://myrocks.io/>
2. Facebook RocksDB. <http://rocksdb.org/>
3. Google LevelDB. <https://github.com/google/leveldb>
4. RocksDB Tuning Guide. <https://github.com/facebook/rocksdb/wiki/RocksDB-Tuning-Guide>
5. The InfluxDB storage engine and the time-structured merge tree (TSM). https://docs.influxdata.com/influxdb/v1.0/concepts/storage_engine/
6. tx-trie 0.18-Succinct Trie Implementation. <https://github.com/hillbig/tx-trie>, 2010.
7. Benoit, D., Demaine, E.D., Munro, J.I., Raman, R., Raman, V., Rao, S.S. Representing trees of higher degree. *Algorithmica* 4, 43 (2005), 275–292.
8. Bingmann, T. STX B+tree C++ Template Classes. <http://idlebox.net/2007/stx-btree/>, 2008.
9. Cooper, B.F., Silberstein, A., Tam, E., Ramakrishnan, R., Sears, R. Benchmarking cloud serving systems with YCSB. In *Proceedings of SOCC'10* (2010), ACM, 143–154.
10. Dong, S. Personal communication, 2017. 2017-08-28.
11. Dong, S., Callaghan, M., Galanis, L., Borthakur, D., Savor, T., Strum, M. Optimizing space amplification in RocksDB. In *Proceedings of CIDR'17*, Volume 3 (2017), 3.
12. Grossi, R., Ottaviano, G. Fast compressed tries through path decompositions. *J. Exp. Algorithm.* 3–4, 19 (2015).
13. Jacobson, G. Space-efficient static trees and graphs. In *Foundations of Computer Science* (1989), IEEE, 549–554.
14. Lakshman, A., Malik, P., Cassandra: A decentralized structured storage system. *ACM SIGOPS Oper. Syst. Rev* 2, 44 (2010), 35–40.
15. Leis, V., Kemper, A., Neumann, T. The adaptive radix tree: ARTful indexing for main-memory databases. In *Proceedings of ICDE'13* (2013), IEEE, 38–49.
16. O'Neil, P., Cheng, E., Gawlick, D., O'Neil, E. The log-structured merge-tree (LSM-tree). *Acta Inform.* 4, 33 (1996), 351–385.
17. Rhea, S., Wang, E., Wong, E., Atkins, E., Storer, N. LittleTable: A time-series database and its uses. In *Proceedings of SIGMOD'17* (2017), ACM, 125–138.
18. Sears, R., Ramakrishnan, R. bLSM: A general purpose log structured merge tree. In *Proceedings of SIGMOD'12* (2012), ACM, 217–228.
19. Zhang, H., Andersen, D.G., Pavlo, A., Kaminsky, M., Ma, L., Shen, R. Reducing the storage overhead of main-memory OLTP databases with hybrid indexes. In *Proceedings of SIGMOD'16* (2016), ACM, 1567–1581.
20. Zhang, H., Lim, H., Leis, V., Andersen, D.G., Kaminsky, M., Keeton, K., Pavlo, A. SuRF: Practical range query filtering with fast succinct tries. In *Proceedings of SIGMOD'18* (2018), ACM, 323–336.
21. Zhou, D., Andersen, D.G., Kaminsky, M. Space-efficient, highperformance rank and select structures on uncompressed bit sequences. In *Proceedings of SEA'13* (2013), Springer, 151–163.

Huanchen Zhang, Hyeontaek Lim, David G. Andersen, and Andrew Pavlo ({huanche1, hl, dga, pavlo}@cs.cmu.edu), Carnegie Mellon University, Pittsburgh, PA, USA.

Viktor Leis (viktor.leis@uni-jena.de), Friedrich Schiller University, Jena, Germany.

Michael Kaminsky (kaminsky@cs.cmu.edu) BrdgAI, Pittsburgh, PA, USA.

Kimberly Keeton (kimberly.keeton@hpe.com), Hewlett Packard Labs, Palo Alto, CA, USA.

© 2021 ACM 0001-0782/21/4 \$15.00

Semantic Web for the Working Ontologist

Effective Modeling for Linked Data, RDFS, and OWL

**Dean Allemang
James Hendler
Fabien Gandon**

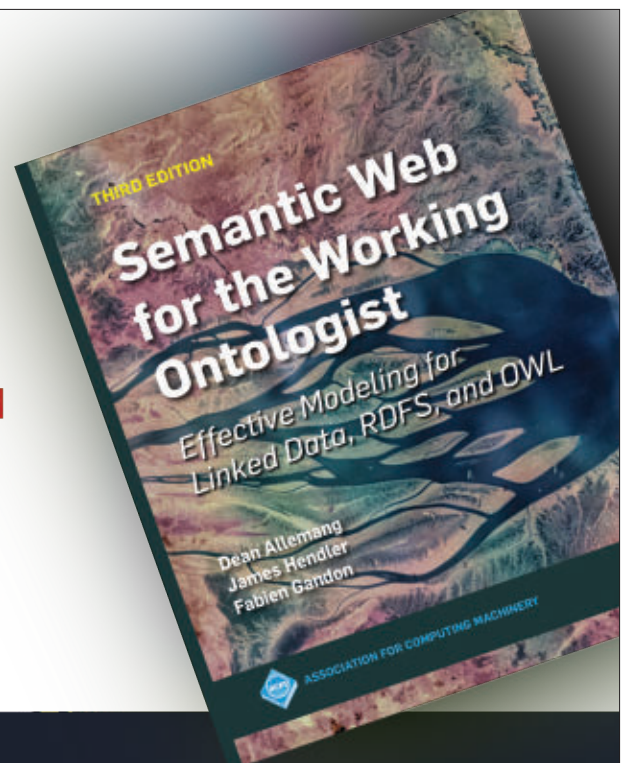
THIRD EDITION

ISBN: 978-1-4503-7617-4

DOI: 10.1145/3382097

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



 **ACM BOOKS**
Collection II

CAREERS

DGIST

Tenure-Track Faculty Public Invitation

It is an honor to have a professor with excellent ability to realize the vision of a 'Convergence University that changes the world with innovation' through convergence education and leading high-tech research along with respect.

1. Positions

Department

Recruitment Field

Emerging Materials Science

- ▶ Materials physics
- ▶ Chemistry

Information and Communication Engineering

- ▶ All areas in electrical engineering and computer science including but not limited to the following
 - Machine learning theory, NLP, computer vision, and other related areas in AI and ML
 - Database / data mining
 - High speed ADC and RF circuit design
 - Neuromorphic devices

Robotics Engineering

- ▶ AI theories and applications for robotics: AI algorithm, deep learning, machine learning, motion planning, robot vision, intelligent control and other related topics
- ▶ Autonomous vehicle technology: computer vision, SLAM, vehicle control, intelligent transportation system and other related research topics
- ▶ General robotics: cooperative robot, industrial robot, humanoid, surgery / rehabilitation robot, exoskeleton, mobile robot and other related research topics
- ▶ All areas in mechanical engineering and electrical engineering related to robotics for exceptional candidates

Energy Science and Engineering

- ▶ All areas in chemistry, physics, materials science, and chemical engineering related to energy conversion, storage, and saving but not limited to the following
 - Design, synthesis, fabrication and modeling of materials and devices
 - Characterization of structure, properties, dynamics and functions
 - Device physics, transmission electron microscope(TEM), spectroscopy

Brain and Cognitive Sciences

- ▶ Cognitive neuroscience
- ▶ Computational neuroscience
- ▶ All areas in cellular/molecular neuroscience for exceptional candidates

New biology

- ▶ Bioinformatics/microbial metagenomics

- ▶ Chemical biology
- ▶ Immunology
- ▶ Plant development and biochemistry

2. Date of Appointment: September 1st, 2021 (Appointment date can be adjusted in consultation with department)

3. Qualification

- ▶ Encourage support for female scientists.
- ▶ With no reasons for disqualification based on related Korean Law(STATE PUBLIC OFFICIALS ACT Article 33)
- ▶ Ph.D. Holder with ability to teach in English required
- ▶ Without distinction of nationality

4. Required Documents

- ▶ DGIST application form
- ▶ 5 representative achievements (Not Passed to submit less than 5 representative achievement)
- ▶ 3 letter of recommendation(Please fill out the list of three recommenders in the application form and submit the recommendation file directly to the department by e-mail if you pass the document review)

5. How to Apply

- ▶ Apply after accessing faculty.dgist.ac.kr website
- ▶ Application Period: Feb 8th, 2021(Mon) ~ Feb 26th, 2021(Fri) 18:00(GMT+09:00)

6. Procedure

Document Screening – Mid March, 2021
Department Interview – Mid April, 2021
Final Interview – Mid May, 2021
Faculty Personnel Committee Review – Mid June, 2021
* Above Schedule is variable depending on DGIST internal situation

7. Matters of Consideration

- ▶ If there is no qualified person, no one can be invited
- ▶ Results of each step will be individually notified via email
- ▶ Appointment shall be canceled in case of
 - 1) a false entry or modification is found in the application form,
 - 2) impossibility to obtain a Ph.D. as of appointment,
 - 3) being rejected in pre-employment medical checkups
- ▶ Other matters not specified in this announcement follow DGIST regulation on Faculty Personnel Management & related regulations
- ▶ Detailed information on contract conditions from DGIST Academic Affairs Team

e-Mail : faculty@dgist.ac.kr

Homepage : faculty.dgist.ac.kr

University of Illinois at Chicago

Open Rank - Multiple Tenure Track Faculty

Located in the heart of Chicago, the UIC CS department is conducting multiple faculty searches this year for multiple tenure track faculty at all ranks. The first is searching broadly within the area of human-computer interaction, including research on a broad range of topics (e.g. mobile, wearable or embedded technologies, ubiquitous computing, robotic interactions, or social and collaborative computing); in a broad range of applications (e.g. healthcare, education, workplace technologies, home or consumer technologies); and broad range of methodologies (e.g. mixed or multi-method, ethnographic, user or case study, time-series analysis, quasi or experimental design). The second is searching for applicants in the area of computational biology to work with UIC's Center for Bioinformatics and Quantitative Biology. The third is searching for applicants from all areas of computer science.

Applications must be submitted at <https://jobs.uic.edu/>, and must include a curriculum vitae, teaching and research statements, and names and addresses of at least three references. Links to a professional website such as Google Scholar or Research Gate are recommended. Applicants may contact the faculty search committee search@cs.uic.edu for more information. For fullest consideration, applications must be submitted by February 11, 2021. Applications will be accepted until the positions are filled.

The department has 40 tenure-system faculty, 4 research faculty, 16 clinical/teaching faculty, and is committed to building a diverse faculty preeminent in its missions of research, teaching, and service to the community. Candidates with experience engaging with a diverse range of faculty, staff, and students, and contributing to a climate of inclusivity are encouraged to discuss their perspectives on these subjects in their application materials.

The University of Illinois at Chicago is an Equal Opportunity, Affirmative Action employer. Minorities, women, veterans and individuals with disabilities are encouraged to apply.

Offers of employment by the University of Illinois may be subject to approval by the University's Board of Trustees and are made contingent upon the candidate's successful completion of any criminal background checks and other pre-employment assessments that may be required for the position being offered. Additional information regarding such pre-employment checks and assessments may be provided as applicable during the hiring process.

The University of Illinois System requires candidates selected for hire to disclose any documented finding of sexual misconduct or sexual harassment and to authorize inquiries to current and former employers regarding findings of sexual misconduct or sexual harassment. For more information, visit <https://www.hr.uillinois.edu/cms/One.aspx?portalId=4292&pageId=1411899>.



DOI:10.1145/3450331

Dennis Shasha

Upstart Puzzles

Roulette Angel

Where will the ball drop when the spinning roulette wheel stops?

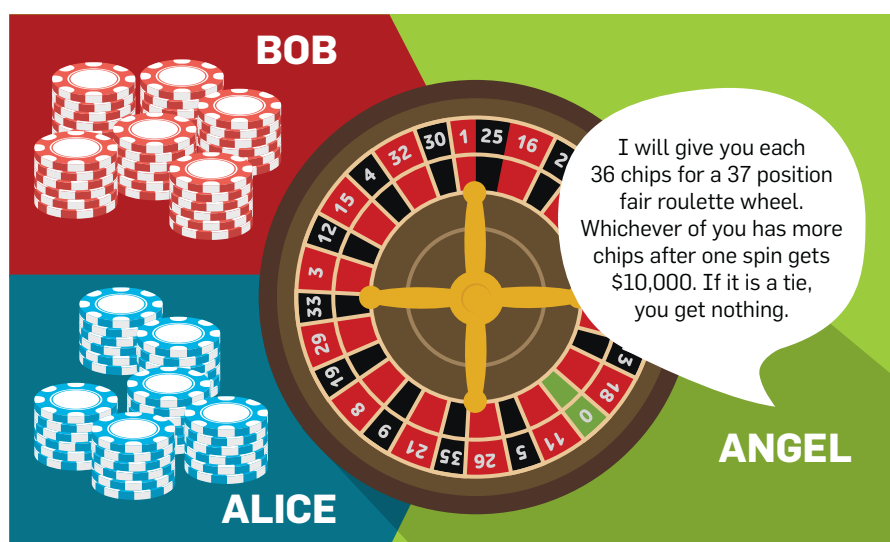
IN THE RULES of roulette, all payouts are as if there are 36 possibilities. So if you bet a single chip on a single number and you win, you receive 36 chips in total (35 plus the one you put in). So, for example, if you put a chip on each of the 36 non-zero numbers and it hit one of those, then you would break even. However, there is also a 0 (and in some casinos a 00). It is possible to get 36 to 1 odds on those, but the point is that there are at least 37 or 38 possibilities. So you have a negative expected value at the game.

Two strangers Bob and Alice receive the following proposition from an eccentric, wealthy, and honest (nearly angelic) benefactor named, well, Angel.

Angel says: There is a fair roulette wheel with a 0 but no 00. I will give you each 36 chips each worth \$1. First Alice will place her bets and then Bob will place his (so he can see Alice's). At the end of the payout from the roll, the two of you have the same number of chips, I disappear and you will never see me again. If one of you has more, that person gets \$10,000. The other gets nothing.

Warm-Up: Alice reasons that all bets have a negative expected value, so she bets nothing. What can Bob do to maximize his chance of winning the \$10,000 and what are his odds?

Solution to Warm-Up: If Bob bets nothing, the result will be a tie, so Bob will receive nothing. Suppose that instead Bob bets on 35 numbers. That is he retains one chip and bets all the others. He has a 35/37 chance of winning on one of his numbers for which he will receive 36 chips. Because he



Sometimes, it is not important to beat the house, just to beat the stranger.

kept one in reserve, he will have 37 chips in total and will win.

OK, now for some challenges for you.

Question: What if Alice had chosen Bob's strategy in the solution to the warm-up, viz. one chip on 35 of the 37 numbers. What could Bob do to maximize his chance of winning?

Solution: Bob should bet on 34 numbers of the 35 numbers on which Alice bets and keep two chips. If the ball falls on any of those 34, he will have 38 chips altogether whereas Alice will have only 37. If the ball falls on any of the numbers that neither he nor Alice picked, he would have two chips, while Alice would have only one. So he would have a 36/37 chance of winning.

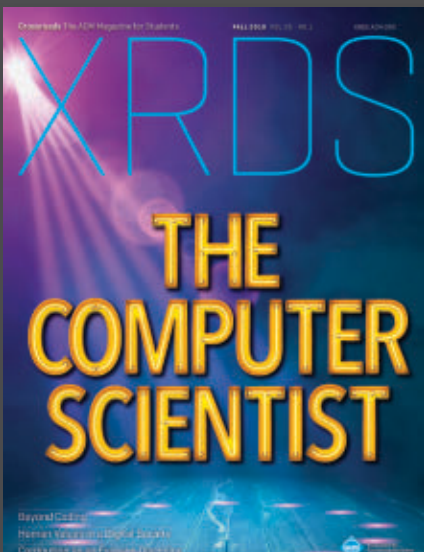
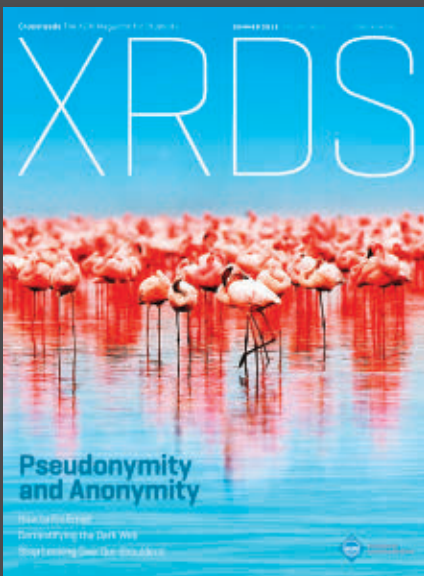
It is clearly a big advantage to go second. So, what if Angel changes the rules a little. Alice still goes first. Bob still sees which numbers Alice chooses, but Bob must bet strictly more chips

than Alice unless Alice bets all 36 in which case Bob also must bet 36.

Question: With these new rules, what is an optimal strategy for Alice? What is an optimal counter-strategy for Bob in that case? What are the odds for each?

Solution: If Alice bets on 35 numbers, then Bob must bet on 36. If the ball rolls on any of Alice's numbers, she wins. Bob can arrange to bet on the two numbers Alice did not bet on and 34 of those that Alice did bet on. So Bob will win 2/37 of the time and Alice 35/37 of the time. If Alice bets on fewer numbers, then her winning probability will go down.

Angel thinks about this and decides this is too favorable to Alice if she can bet more than 18 and too favorable to Bob if he sees what she bets. So, Angel limits Alice to bet 18 or fewer. [CONTINUED ON P. 175]



XRDS

At *XRDS*, our mission is to empower computer science students around the world. We deliver high-quality content that makes the complexity and diversity of this ever-evolving field accessible. We are a student magazine run by students, for students, which gives us a unique opportunity to share our voices and shape the future leaders of our field.

Accessible, High-Quality, In-Depth Content We are dedicated to making cutting-edge research within the broader field of computer science accessible to students of all levels. We bring fresh perspectives on core topics, adding socially and culturally relevant dimensions to the lessons learned in the classroom.

Independently Run by Students *XRDS* is run as a student venture within the ACM by a diverse and inclusive team of engaged student volunteers from all over the world. We have the privilege and the responsibility of representing diverse and critical perspectives on computing technology. Our independence and willingness to take risks make us truly unique as a magazine. This serves as our guide for the topics we pursue and in the editorial positions that we take.

Supporting and Connecting Students At *XRDS*, our goal is to help students reach their potential by providing access to resources and connecting them to the global computer science community. Through our content, we help students deepen their understanding of the field, advance their education and careers, and become better citizens within their respective communities.

XRDS is the flagship magazine for student members of the Association for Computing Machinery [ACM].

www.xrds.acm.org



Association for
Computing Machinery



SIGGRAPH 2021

THE PREMIER **CONFERENCE & EXHIBITION** IN
COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES



SAVE THE DATE

9-13 AUGUST

Join us at SIGGRAPH 2021 to celebrate and honor the past, present, and future of computer graphics and interactive techniques. Celebrating 48 years of excellence, SIGGRAPH presents a high-quality experience showcasing the latest research, cutting-edge ideas, and breakthrough discoveries.

[S2021.SIGGRAPH.ORG](https://s2021.siggraph.org)



Sponsored by
ACM **SIGGRAPH**

THE 48TH INTERNATIONAL CONFERENCE & EXHIBITION
ON COMPUTER GRAPHICS & INTERACTIVE TECHNIQUES