

# COMMUNICATIONS

CACM.ACM.ORG OF THE ACM 11/2020 VOL.63 NO.11

## Special Section on Latin America Region



Coding at a Crossroads

Reason-Checking Fake News

Terahertz Networks Move Closer to Reality

What Should Be Done About Social Media?



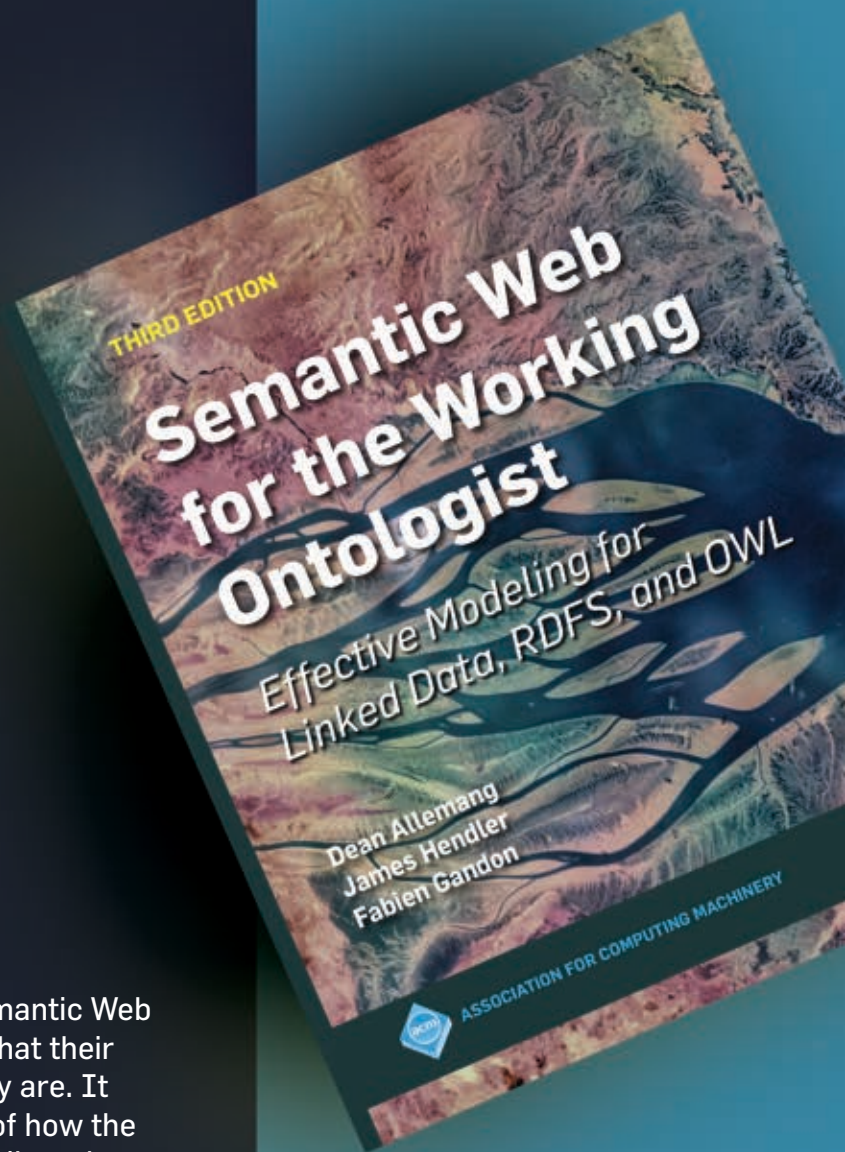
## ACM BOOKS Collection II

Enterprises have made amazing advances by taking advantage of data about their business to provide predictions and understanding of their customers, markets, and products. But as the world of business becomes more interconnected and global, enterprise data is no long a monolith; it is just a part of a vast web of data. Managing data on a world-wide scale is a key capability for any business today.

The Semantic Web treats data as a distributed resource on the scale of the World Wide Web, and incorporates features to address the challenges of massive data distribution as part of its basic design. The aim of the first two editions was to motivate the Semantic Web technology stack from end-to-end; to describe not only what the Semantic Web standards are and how they work, but also what their goals are and why they were designed as they are. It tells a coherent story from beginning to end of how the standards work to manage a world-wide distributed web of knowledge in a meaningful way.

The third edition builds on this foundation to bring Semantic Web practice to enterprise. Fabien Gandon joins Dean Allemang and Jim Hendler, bringing with him years of experience in global linked data, to open up the story to a modern view of global linked data. While the overall story is the same, the examples have been brought up to date and applied in a modern setting, where enterprise and global data come together as a living, linked network of data. Also included with the third edition, all of the data sets and queries are available online for study and experimentation at: [data.world/swwo](http://data.world/swwo).

<http://books.acm.org>  
<http://store.morganclaypool.com/acm>



**Semantic Web for the  
Working Ontologist**  
*Effective Modeling  
for Linked Data, RDFS,  
and OWL*

**THIRD EDITION**

**Dean Allemang  
James Hendler  
Fabien Gandon**

ISBN: 978-1-4503-7617-4  
DOI: 10.1145/3382097



# EICS 2021

The 13th ACM SIGCHI Symposium on  
Engineering Interactive Computing Systems



Eindhoven, The Netherlands

8-11 June 2021



## Interactive systems

For more info visit  
[eics.acm.org/eics2021](https://eics.acm.org/eics2021)

Modelling and analysis, interaction design, software architectures and specification, frameworks, toolkits, APIs, formal methods, engineering innovative applications, engineering hardware/software integration or users' experiences and activities.

Sponsors



## Departments

- 5 **Vardi's Insights**  
**What Should Be Done About Social Media?**  
*By Moshe Y. Vardi*
- 
- 7 **Career Paths in Computing**  
**A Career Unfolds in Phases**  
*By Celeste M. Rohlifing*
- 
- 8 **Letters to the Editor**  
**Weighing Grad School Payback**
- 
- 12 **BLOG@CACM**  
**Bringing Industry Back to Conferences, and Paying for Results**  
David Patterson wants to boost industry submissions to conferences, while Yegor Bugayenko suggests productivity should govern coders' pay when they work from home.
- 
- 155 **Careers**

## Last Byte

- 160 **Q&A**  
**Tackling the Challenges of CS Education**  
Chris Stephenson on the complex challenges that continue to plague the computer science education community.  
*By Leah Hoffmann*

## News



- 14 **Seeking Artificial Common Sense**  
The long-standing goal of providing artificial intelligence some measure of common sense remains elusive.  
*By Don Monroe*
- 
- 17 **Natural Language Misunderstanding**  
How do we eliminate bias in automated speech recognition?  
*By Keith Kirkpatrick*
- 
- 19 **Terahertz Networks Move Closer to Reality**  
The desire for faster, higher-frequency wireless networking is a constant. Terahertz technology could deliver large gains.  
*By Samuel Greengard*

## Viewpoints

- 22 **Privacy**  
**Digital Contact Tracing May Protect Privacy, But It Is Unlikely to Stop the Pandemic**  
Considering the potential benefits versus the risks of privacy-enhancing technologies.  
*By Lorrie Faith Cranor*
- 
- 25 **Legally Speaking**  
**Copyright's Online Service Providers Safe Harbors Under Siege**  
Reviewing the most significant changes recommended in the recently released U.S. Copyright Office Section 512 Study.  
*By Pamela Samuelson*
- 
- 28 **Economic and Business Dimensions**  
**Using Data and Respecting Users**  
Three technical and legal approaches that create value from data and foster user trust.  
*By Marshall W. Van Alstyne and Alisa Lenart*
- 
- 31 **Education**  
**It Is Time for More Critical CS Education**  
By which 'critical' means an intellectual stance of skepticism, centering the consequences, limitations, and unjust impacts of computing in society.  
*By Amy J. Ko et al.*
- 
- 34 **Viewpoint**  
**Where Should Your IT Constraint Be? The Case of the Financial Services Industry**  
Locating the strategic location of the IT function constraint.  
*By Boaz Ronen and Alex Coman*
- 
- 38 **Viewpoint**  
**Reason-Checking Fake News**  
Using argument technology to strengthen critical literacy skills for assessing media reports.  
*By Jacky Visser, John Lawrence, and Chris Reed*





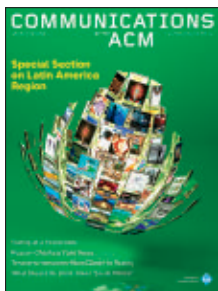
Special Section



42 **Latin America Region**  
This issue's special section spotlights the countries that embody Latin America—a striving and thriving region of technological innovation. The contributors, representing such countries as Argentina, Brazil, Chile, Colombia, Mexico, Peru, Uruguay, and more, share some of the latest trends and technical advances from LATAM.



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



**About the Cover:**  
Cover illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: BendableSound image by Centro de Investigación Científica y de Educación Superior de Ensenada. LIDAR Amazon rainforest image courtesy of CADAF Project, LMF/INPA (Brazil). Brain image courtesy of the Imaging Sciences Laboratory/imaglabs.org. Sao Paulo street traffic image by LucVi/Shutterstock.com; Machu Picchu photo by Bob Pool/Shutterstock.com; Bike Station photo by Cassiohabib/Shutterstock.com; Molina Healthcare photo by rafapress/Shutterstock.com; Mexico City mall photo by irvingthewaffle/Shutterstock.com. Additional stock images from Shutterstock.com.

Practice



108 **Five Nonobvious Remote Work Techniques**  
Emulating the efficiency of in-person conversations.  
*By Thomas A. Limoncelli*

111 **Data on the Outside versus Data on the Inside**  
Data kept outside SQL has different characteristics from data kept inside.  
*By Pat Helland*

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

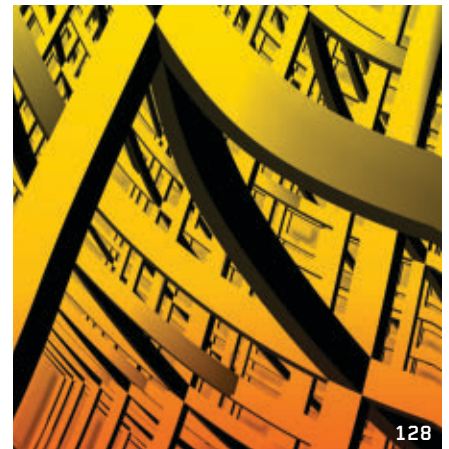
Contributed Articles

120 **Coding at a Crossroads**  
While millions of students worldwide have enjoyed coding experiences over the last decade, the next challenge is spreading educational values and approaches.  
*By Mitchel Resnick and Natalie Rusk*



Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/coding-at-a-crossroads>

Review Articles



128 **The Graph Isomorphism Problem**  
Exploring the theoretical and practical aspects of the graph isomorphism problem.  
*By Martin Grohe and Pascal Schweitzer*

Research Highlights

138 **Technical Perspective**  
**When the Adversary Is Your Friend**  
*By Alexei A. Efros and Aaron Hertzmann*

139 **Generative Adversarial Networks**  
*By Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio*

145 **Technical Perspective**  
**BLeak: Semantics-Aware Leak Detection in the Web**  
*By Harry Xu*

146 **BLeak: Automatically Debugging Memory Leaks in Web Applications**  
*By John Vilks and Emery D. Berger*



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**

Andrew Rosenbloom

**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**

acmhhelp@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, 10<sup>th</sup> 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, 10<sup>th</sup> 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**

Terry Benzel; Michael L. Best; Judith Bishop;  
Lorrie Cranor; Boi Falting; James Gimmelmann;  
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; Jeff Chase;  
Premkumar T. Devanbu; Jane Cleland-Huang;  
Yannis Ioannidis; Trent Jaeger; Somesh Jha;  
Gal A. Kaminka; Ben C. Lee; Igor Markov;  
Lionel M. Ni; Doina Precup; Shankar Sastry;  
m.c. schraefel; Ron Shamir; 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 Efros; Bryan Ford; Alon Halevy;  
Gernot Heiser; Takeo Igarashi;  
Srinivasan Keshav; Sven Koenig;  
Ran Libeskind-Hadas; Karen Liu; Greg Morrisett;  
Tim Roughgarden; Guy Steele, Jr.;  
Robert Williamson; Margaret H. Wright;  
Nicolai Zeldovich; Andreas Zeller

**SPECIAL SECTIONS**

**Co-Chairs**

Sriram Rajamani, Jakob Rehof, and Haibo Chen

**Board Members**

Sue Moon; P.J. Narayana; Tao Xie;  
Kenjiro Taura; David Padua

**ACM Copyright Notice**

Copyright © 2020 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 acmhhelp@acm.org.

**COMMUNICATIONS OF THE ACM**

(ISSN 0001-0782) is published monthly by ACM Media, 1601 Broadway, 10<sup>th</sup> 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, 10<sup>th</sup> Floor New York, NY 10019-7434 USA

Printed in the USA.



Association for Computing Machinery







Moshe Y. Vardi

DOI:10.1145/3424762

# What Should Be Done About Social Media?

ONE OF THE most basic and urgent policy questions is how to tackle the rising role of social media in our public sphere. As social media has proliferated across the globe, societies have had to grapple with its implications for both exercising and constraining speech. While social media has provided a platform for countless individuals to express their opinions, many argue that social-media companies must adopt more accountability for harmful content published on their sites.

To illustrate the salience of technological change in the world of social media, the case of Facebook is timely. Over the last several years, Facebook has been involved in a long series of controversial issues, from Cambridge Analytica to hate speech in Myanmar. Responding to the bad publicity that accompanied these disclosures, Facebook's CEO Mark Zuckerberg wrote a *Washington Post* op-ed in 2019 calling for increasing regulation of the Internet in four areas: harmful content, election protection, effective privacy and data protection, and data portability.

Until 2014, Facebook's motto was, "Move fast and break things. Unless you are breaking stuff, you are not moving fast enough." Despite the benefits of innovation and change, "breaking things" can have profound and dangerous unintended societal consequences. The case that social media has become an instrument for undermining democracy is a strong one. It is now widely accepted that social media seriously affected the 2016 Brexit referendum in Britain and the presidential election in the U.S. It is this cavalier attitude about breaking things that led *Wall Street Journal* columnist Peggy Noonan to de-

scribe Silicon Valley executives as "moral Martians."

A discussion of modern Internet regulation merits mention of Section 230 of the Communications Decency Act of 1996, a fundamental piece of U.S. legislation that provides immunity from liability for providers and users of an "interactive computer service" who publish information provided by third-party users. By allowing Facebook and other Internet companies to operate as a platform, rather than as a publisher, Section 230 frees them from liability for the content they publish. The explosive growth of social-media platforms would have not been possible without Section 230. At the same time, it is doubtful Congress in 1996 could have conceptualized anything similar to social media. One can also argue that Facebook is quite far from being a neutral platform because of its algorithm-based system that generates content based on users' preferences. In fact, the proliferation of "bad speech" on social-media platforms has become politically untenable, and now all social-media platforms are actively fighting "bad speech." Thus, in spite of Section 230, social-media platforms seem to be accepting some responsibility for the content they publish. In other words, they are starting to behave with some restraint, like publishers, rather than platforms.

It is not at all clear, however, whether a platform like Facebook, which also owns Instagram and WhatsApp, with more than 2.5 billion active users, can behave like a traditional publisher. First, there is the difficulty of vetting content from a very large number of users. With just over 50,000 employees, Facebook clearly cannot have people review all its content; algorithmic filtering is a must. But, if we have learned

anything over the last few years, it is how good people are at outsmarting algorithms. More fundamentally, however, do we really want Facebook to regulate the speech of more than 2.5 billion people? No government in the world has such power to regulate the speech of almost a third of humanity. Of course, traditional publishers regulate speech on their platforms, but there is a multiplicity of such outlets with no single authority having a monopoly on deciding for or against certain content. In contrast, there is only *one* Facebook.

The basic policy question—how to regulate speech on social media platforms—seems inseparable from another policy concern, namely how to deal with the concentration of power in technology. The five largest U.S. corporations are all tech companies—Alphabet, Amazon, Apple, Facebook, and Microsoft—with combined market capitalization approaching seven trillion dollars. For this reason, the tech sector is often called "Big Tech" these days. In a 2018 book, *The Curse of Bigness: Antitrust in the New Gilded Age*, legal scholar Tim Wu argues the U.S. must enforce anti-trust laws against such corporations.

Breaking well-integrated corporations is difficult, but there are some easier options on the table. Should Facebook be forced to spin off Instagram and WhatsApp? Should Google be forced to spin off YouTube? The time has come to put these questions on the table!

Follow me on Facebook and Twitter. 

**Moshe Y. Vardi** ([vardi@cs.rice.edu](mailto:vardi@cs.rice.edu)) is the Karen Ostrum George Distinguished Service Professor in Computational Engineering and Director of the Ken Kennedy Institute for Information Technology at Rice University, Houston, TX, USA. He is the former Editor-in-Chief of *Communications*.

Copyright held by author.

# Publish Your Work Open Access With ACM!

ACM offers a variety of Open Access publishing options to ensure that your work is disseminated to the widest possible readership of computer scientists around the world.



Please visit ACM's website to learn more about ACM's innovative approach to Open Access at:  
<https://www.acm.org/openaccess>



Association for  
Computing Machinery





# CAREER PATHS IN COMPUTING

DOI:10.1145/3422082

## Computing enabled me to . . . **A Career Unfolds in Phases**



### NAME

**Celeste M. Rohlfing**

### BACKGROUND

**Born in Bryn Mawr, PA;  
now living in Durham, NC**

### CURRENT JOB TITLE/EMPLOYER

**Retired from National  
Science Foundation and  
American Association for  
the Advancement of Science**

### EDUCATION

**Ph.D. Chemistry,  
Princeton University, 1983**

**L**OOKING BACK, I can divide my career into three phases: practicing science, enabling science, and advocating for science. As an undergraduate, I was passionate about chemistry and mathematics, but in my junior year, I discovered how the power of computing could address research questions in chemistry. It comes as no surprise that I pursued a Ph.D. in theoretical and computational chemistry. Wanting access to the world's fastest supercomputers (Cray, anyone?), I took my first job at a U.S. Department of Energy national laboratory.

By age 38, I was eager to make the transition to the second phase of my


career. Although I loved the intellectual stimulation of conducting research, I perceived this career path as a long, straight line stretching into the future. Ultimately, I recognized it would not satisfy my internal yearning to *do something more*, to have an impact beyond my scientific publications.

In a leap of faith, I moved my family across the country and started a new career at the U.S. National Science Foundation. This proved to be the environment where I could do that undefined *something more*. I created a cyber-enabled chemical sciences program in the early 2000s and guided funding initiatives in nanotechnology and quantum information science. Eventually, I became a senior executive who enabled science through oversight of a \$1.5B portfolio of projects across chemistry, materials science, mathematics, physics, and astronomy.

After President Obama's election, I received an invitation to be Assistant Director for Physical Sciences in the White House Office of Science and Technology Policy. Although this wasn't a change I was actively seeking, I took a leave from the National Science Foundation to immerse myself in the world of science policy and explore other ways to contribute to the advancement of science. After the 2013 shutdown of the Federal Government, I was frustrated because I could neither publicly promote increased budgets for my agency nor ask others to do so. At age 57, I was ready to commit to advocating for science. So, I became Chief Operating Officer at the American Association for the Advancement of Science, a multidisciplinary

scientific society and publisher of the journal *Science*.

In this position, I was able to speak out against the restriction on using federal dollars for public health studies on gun violence; share the latest research findings on the detrimental effects of implicit bias in peer review journal articles with scientific society publishers; and advise decision-makers on the role of science in policymaking. I also witnessed the growing politicization of science, including the vocalization of an anti-science mentality from various elected leaders, culminating in my organization's critical support for the first-ever "March for Science." Neglect, distrust, and willful disrespect of scientific expertise are horrifyingly evident in the U.S. right now as the mortality rate from Covid-19 climbs. Alas, this is a predictable result of a decades-long assault on scientific expertise.

From my perspective of a long career in science, I can confidently state that our government at local, state, and federal levels cannot function properly without the input of evidence and expertise to policy making. If science were routinely at the table, what different policy solutions might have arisen by now for a host of societal challenges in education, economy, health, housing, energy, and environment? I challenge you to get involved in both science policy and science *for* policy, by offering potential solutions from research and applications in areas such as data science, robotics, and artificial intelligence. It is vital that you defend science from attacks, celebrate its value to society, and, please, *do something more*. 

DOI:10.1145/3425745

# Weighing Grad School Payback

**I**N HIS SEPTEMBER 2020 column (p. 5), Moshe Vardi rightly criticizes the Trump Administration’s policies prohibiting foreign graduate students and points out the dearth of domestic graduate students willing to fill positions. He asks to understand the root of this problem, and I would like to share my perspective on why I chose another path.

During my undergraduate program, a few professors encouraged me to consider graduate school. They said it had been the best decision they had made. However, I could not help but feel that I was talking to lottery winners about buying tickets. To me, graduate school looked like a long, sloggy prospect with an uncertain outcome. The conventional wisdom was that graduate students were the grunts of the academic workforce, suffered disproportionate grievances, and won disproportionate low rewards. The stories of destitute adjunct faculty in academia pointed to a future with great risk.

For me at least, not going to graduate school was a question of incentives. To convince students to attend graduate school, they need to be convinced their lives will be better by going.

**John Boyd**, Brooklyn, NY, USA

**Author’s response:**

*John Boyd’s comment on the strict winnowing process that leads to tenured positions in major research universities is fair, but it cannot be the full explanation. Most doctorate holders in computer science, after all, end up in industry and not in academia. Also, other fields, such as music or sport, also have a strict winnowing process, yet they do not lack applicants.*

**Moshe Y. Vardi**, Houston, TX, USA

**Editor-in-Chief’s response:**

*These are valid perspectives on education—its utility in acquiring wealth and securing career opportunities. Another classical view of education is in*

*its reward in both the journey, and the growth that comes from the journey that enables us to appreciate life, science, technology, and those who both pursue and remarkably, shape its trajectory! I recommend Nick Feamster’s thoughtful piece on “Do You Need a Ph.D.?” on Medium (<https://bit.ly/35P4xie>).*

**Andrew A. Chien**, Chicago, IL, USA

**Body of Evidence**

I read Vinton G. Cerf’s August 2020 column (as I do all of them) with interest. Before folks adopt “Internet of Medical Things,” it might be useful to note that a different term has been in circulation for a few years, one that encompasses medical things and other wearables. A mathematician colleague of mine at RAND Corporation, Mary Lee, was one of the early commenters on the label “Internet of Bodies” or IOB (see <https://wapo.st/2PIB54s>).

I will admit it took me a while to get used to the IOB label, but it is briefer and, in my opinion, more flexible than IOMT. Moreover, it evokes some new possibilities for, say, peer-to-peer networking as well as the “phone home” aspects of most of today’s networked medical things. Mary Lee has a report on this topic in press that surveys the landscape and related issues, so stay tuned for that.

**Marjory S. Blumenthal**, Washington, D.C., USA

There are three fundamental assumptions toward the success of the Internet of Medical Things proposed by Vinton G. Cerf in his August column (p. 5). The first assumption is that sensory technologies are capable of continuously measuring a variety of aspects of human physiology. The second assumption relies on achieving ultimate accuracy by minimizing false positive and negative detection rates of both diagnosing onset of a condition and helping monitor existing conditions. The third assumption relies on the consent of individuals, either those who are disease-free or

**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](mailto:acmm mediasales@acm.org)





live with one or more chronic conditions, to allow the use of technologies to continuously analyze and report on the functionality of their organs, often to entities other than their trusted medical personnel such as insurance companies.

A fourth fundamental assumption not mentioned in Cerf's column is the inability of sensory technologies to accurately diagnose the majority of life-threatening conditions. It is indeed possible to assess controlled conditions such as hypertension and type 2 diabetes and even more complex conditions, for example, monitoring heart failure by incorporating CardioMEMS's wireless sensors implanted in patients' pulmonary arteries.<sup>1</sup> Conditions such as cancer, heart disease, and bone disease, however, often require the use of technologies available only in hospitals, considering their high prices and large sizes. To diagnose heart conditions, patients may be evaluated by stress tests, cardiac computerized tomographies, and echocardiograms, among more invasive tests such as coronary angiograms and myocardial biopsies. By asking patients about their symptoms, a physician can diagnose unpredictable conditions such as spontaneous pneumothorax (collapsed lung without any reason in a healthy individual), which results in a sharp and continuous chest pain and leaves no option but to visit the emergency room, but ultimately the diagnosis is confirmed by an X-ray, a technology not yet available as a wearable.<sup>2</sup>

Mainstream adoption of the Internet of Medical Things would require addressing all four assumptions. I believe bringing hospital-only sensory technologies to individuals as wearables or even as technologies not yet existent is possible given incredible progress in combining advances in health care and computer science. Indeed, it currently seems unrealistic for a complex medical device to be used in a nonhospital setting, for example, acquiring MRI- or bone-density-like scans at home, even multiple times a day. However, one angle to view sensory technologies of the Internet of Medical Things at their current stage is to compare the current adoption and usefulness of the Inter-

net to that when it was invented many decades ago.

#### References

1. Abbott. CardioMEMS™ HF System. Clinical compendium (Feb. 2020).
2. Kartoun, U. Improving the management of spontaneous pneumothorax. *European Respiratory J.* 52, 6 (Dec. 2018), 1–3.

Uri Kartoun, Cambridge, MA, USA

#### A Letter Apart

In the June 2020 issue (p. 6), Donald Costello suggests ACM be renamed ACP for Association of Computing Professionals.

While having an almost-as-long association with ACM as Costello, I agree that 'Machinery' is a term that is past its best-by date, but the proposed change is too acute.

Essentially it would change the Association from being about computing to being about people who have something to do with computing. There are surely sufficient social media groups and forums for those people.

By all means change the 'M' (perhaps to 'E' for Engineering, or 'K' for Knowledge, or almost any other letter), but surely ACM's strength and advantage has always been that it has focused on computing machinery, software, and algorithms—and everything those have achieved—rather than the people.

Mike Cowlshaw, Coventry, England, U.K.

#### Erratum

The article by Fay Cobb Payton and Alexa Busch "Examining Undergraduate Computer Science Participation in North Carolina" (Aug 2020, p. 60) contained labeling errors within the figures in the print edition. The lines labeled for Duke and UNCC were inadvertently switched. *Communications* regrets this error.

*Communications* welcomes your opinion. To submit a Letter to the Editor, please limit yourself to 500 words or less, and send to [letters@cacm.acm.org](mailto:letters@cacm.acm.org).

© 2020 ACM 0001-0782/20/11 \$15.00

Coming Next Month in **COMMUNICATIONS**

**Green AI**

**The Dark Triad and Insider Threats in Cyber Security**

**Measuring Internet Speed**

**The Life of a Data Byte**

**Security Analysis of SMS as a Second Factor of Authentication**

**XNOR Net**

**Operationalizing AI Ethics Principles**

**Silicon Politics**

Plus the latest news about artificial skin, technology for the visually impaired, and contact tracing while preserving privacy.

# 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](http://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](http://www.acm.org/code-of-ethics)) and ACM's Policy Against Harassment ([www.acm.org/about-acm/policy-against-harassment](http://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](mailto:acmhelp@acm.org)  
[www.acm.org/join/CAPP](http://www.acm.org/join/CAPP)



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.



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

DOI:10.1145/3422628

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

## Bringing Industry Back to Conferences, and Paying for Results

*David Patterson wants to boost industry submissions to conferences, while Yegor Bugayenko suggests productivity should govern coders' pay when they work from home.*



**David Patterson**  
Restoring Industry Participation in Computer Science Conferences

<https://bit.ly/3LLz4ml>

July 17, 2020

Computer Science has had a long, friendly, synergistic relationship between industry and academia. For example, when I started going to the International Symposium on Computer Architecture (ISCA, <https://iscaconf.org/isca2020/>) in the 1970s, 40% of the papers were on real products from industry.

Industry papers fell from 40% in the 1970s to 10% recently at ISCA (<https://www.sigarch.org/publication-trends-at-isca/>), and even that 10% includes papers based more on industrial research than on industrial products. If these trends continue, the historic bond between computer architecture research and practice could fade, making it harder to understand the problems facing industry and for our research to have impact. When I complained

at ISCA 2019 about the lack of papers on real industrial products, ACM SIGARCH chair Sarita Adve assigned me to help fix the problem for 2020.

I thought the only hope was a separate submission process with a separate program committee (PC) whose members believed retrospective papers on industrial products were valuable complements to academic research papers. The PC members also needed to understand company concerns about patent issues or trade secrets may mean some details are not revealed. We tried the following guidelines for industry papers, which differ from regular ISCA and other conferences that have industry tracks:

1. **Recruit submissions.** Few product architects have permission—to let alone time or motivation—to write papers. Given the paucity of product papers, we reached out to nine companies with our top 10 reasons to publish at ISCA (<https://bit.ly/3lNuPqY>).

2. **No intern/sabbatical papers.** Papers by visiting students or faculty already appear at ISCA. We required the

first, and virtually all, authors must work in industry.

3. **Real product papers.** We selected papers based on real products, not prototypes developed in industry research labs.

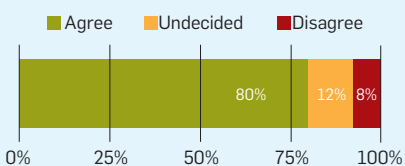
4. **Identify products and authors.** ISCA uses double-blind reviewing, but industry paper reviewers want to know the actual name of the company and the product and whether the authors are its architects.

5. **Later submission deadline.** Management must approve industry papers before submission (often involving multiple rounds of redaction), and there can be restrictions about filing patents before submitting a paper. We delayed the paper deadline two months to improve the odds of receiving papers. Besides, reviewing  $\approx 20$  papers takes less time than  $\approx 400$ .

6. **Small PC.** We limited the PC to nine people (<https://bit.ly/3bvOBIV>) to facilitate full discussions of the papers at the PC meeting.

The competition was stiff for the 19 papers we received (three recruited), with the majority of papers being strong. They were so strong that Sophia Shao and I are guest-editing a special issue of *IEEE Micro* (<https://bit.ly/32TqV77>) for the papers we could not accept. PC members also on the regular ISCA PC found the industry papers very engaging and brought interesting perspectives with different values than the typical ISCA submission. We accepted 5/19 or 26%, only one was recruited (the regular ISCA PC accepted 77/428, or 18%). Each paper proudly bore the label “Industrial Product.”

### The industry track session makes the ISCA program stronger/more exciting.



I was delighted the papers appeared in the honored first session of ISCA. The big question was the audience reaction. The graph shows the main survey question about the industry track (132 responses). I am not sure what other questions would get 80% agreement from skeptical computer architects—free alcohol and ice cream at the ISCA reception? Given the many changes we tried, I was thrilled with these results.

I suspect few of the selected papers would have survived the traditional ISCA process, as reviewers would expect different content and evaluations, so I believe a separate PC is critical to the success of an industry session. I think restricting papers to be on real products by actual industry authors is what made attendees say the ISCA program was stronger and more exciting than in the recent past. The small PC led to more thorough discussions than a typical PC meeting, and the delayed deadline increased our submissions. We intend to follow all six guidelines in the future.

Some conferences already have a version of an industry track (<https://bit.ly/2Gt5qCD>). Given the

- ▶ importance of a good ties between industry and research in computer science,
  - ▶ decline in participation by industry in many conferences, and
  - ▶ strong positive reaction from both authors and the audience to this experiment,
- perhaps more conferences should institute an industry track?

### Comments

*Thanks David, for your article on this important topic. I agree with all you say (a standard I don't achieve with even my own writings). From my experience, having been an ACM member since 1973, the connection between academe and industry has monotonically declined to a very sad state. I applaud efforts to reverse the trend; yours being an excellent role model that I hope will inspire others.*

—Robert Akscyn



### Yegor Bugayenko The Remote Revolution Has to be Driven by Output, Not Salaries

<https://bit.ly/2XmoqBL>

May 29, 2020

Across the world, millions of people are giving remote work a go. The COVID-19 pandemic has thrown us all for a loop and forced countless companies to shutter offices, warehouses, and everything else.

Before the pandemic, remote work was already gaining steam in many industries and circles. For workers, it is promised as “liberation” or “freedom” or whatever buzzword makes people feel better.

Some companies see remote as a panacea, too. If everyone’s working remotely, there’s less chance of office politics, right? Costs go down as well because you don’t have to rent, heat, and furnish office space. Blah, blah, blah.

Many companies will realize how difficult remote work is. Monitoring employees remotely is hard. Keeping teams motivated and on task is even more difficult. Maintaining productivity is next to impossible, especially if you don’t adapt your management and compensation models for the reality of remote life.

Working remotely works best when you pay people for results, not by the hour. Many companies are still paying their employees salaries, and they will find out just how hard it is to motivate an off-site employee when they’re paid the same no matter their output.

Think about it; if you will get paid the same whether you code for eight hours or watch Netflix all day, which are you going to do? Sure, a lot of employees will try to be productive; they’ll wake up, work for an hour or two, then take a break.

A 20-minute break can quickly turn into two hours. Some employees will remain productive, and holding people accountable can keep people on task, but as long as the rewards remain the same, productivity will slowly slip.

It is harder to control employees remotely. You can’t walk around the office to make sure people are on task and not goofing off. Nor can you swing by the desk to check in and ask for an update.

Need to schedule a team meeting? It’s easier to gather the team when they’re all in an office rather than scattered across town. Software problems, hardware issues, scheduling conflicts,

and all that become more frequent.

So what’s the answer? If people are being paid to deliver concrete, measurable results, a lot of these problems will disappear, or at least be less of a hassle. Don’t pay by the hour, pay by the output.

For one, you don’t have to monitor people if they’re only being paid for production. If a worker sat around watching Netflix all day and didn’t get his work done, that’s not just your problem; it’s also his problem come payday.

People have more incentive to make certain meeting software is installed correctly and their computer is set up properly before a meeting starts. When you pay people by the hour, they get paid even while they installing updates and turning on their mic. When they’re compensated for results, they’re wasting not only your time, but also their own.

Some predict the COVID-19 pandemic will encourage more companies to shift away from traditional office spaces. Working from home or the coffee shop will become the norm because people will realize how wonderful it is.

What’s more likely is that companies will realize how much of a hassle remote work is, and how out-of-tune it is with hourly pay and fixed salaries. Employees may find themselves more worried (rather than less) about their bosses looking over their shoulders or bugging them for updates.

Many companies that had given the work-from-home revolution a go before the pandemic have already scaled back their efforts. Google, Best Buy, and Yahoo are just a few that have found out how tough managing operations remotely is. Many more will learn painful lessons in the months ahead.

If you want to go ahead with the remote revolution, then switch to results-based payment models. Pay people and pay them well when they deliver credible results and outputs. This will provide employees with the motivation to succeed on their own.

---

**David Patterson** is a Google distinguished engineer; a professor for the University of California, Berkeley; RISC-V International Vice Chair, and the RISC-V International Open Source (RIOS) Laboratory Director. His newest book is *The RISC-V Reader: An Open Architecture Atlas* and his best known is *Computer Architecture: A Quantitative Approach*. He and his co-author John Hennessy shared the 2017 ACM A.M. Turing Award. **Yegor Bugayenko** is founder and CEO of software engineering and management platform Zerocracy.

© 2020 ACM 0001-0782/20/11 \$15.00

## Seeking Artificial Common Sense

*The long-standing goal of providing artificial intelligence some measure of common sense remains elusive.*

**A**LTHOUGH ARTIFICIAL INTELLIGENCE (AI) has made great strides in recent years, it still struggles to provide useful guidance about unstructured events in the physical or social world. In short, computer programs lack common sense.

“Think of it as the tens of millions of rules of thumb about how the world works that are almost never explicitly communicated,” said Doug Lenat of Cycorp, in Austin, TX. Beyond these implicit rules, though, commonsense systems need to make proper deductions from them and from other, explicit statements, he said. “If you are unable to do logical reasoning, then you don’t have common sense.”

This combination is still largely unrealized; in spite of impressive recent successes of machine learning in extracting patterns from massive data sets of speech and images, they often fail in ways that reveal their shallow “understanding.” Nonetheless, many researchers suspect hybrid systems that combine statistical techniques with more formal methods could approach common sense.

Importantly, such systems could also genuinely describe how they came

to a conclusion, creating true “explainable AI” (see “AI, Explain Yourself,” *Communications* 61, 11, Nov. 2018).

### Elusive Metrics

It can be difficult to determine whether a system really has common sense, or is just faking it. “That’s a classic issue with how AI is perceived versus what it’s doing,” said David Ferrucci of Elemental Cognition in Wilton, CT, who previously led IBM’s Watson project,

the results of which displayed superhuman performance on the television quiz show *Jeopardy!* “One of the problems with AI is that we project,” Ferrucci said, offering the example of thinking that, because human contestants understood what they read, Watson must understand, too. “That’s not the way it’s solving the problem.” he said.

Because tasks and assessments of common sense often are formulated linguistically, they become “tied up

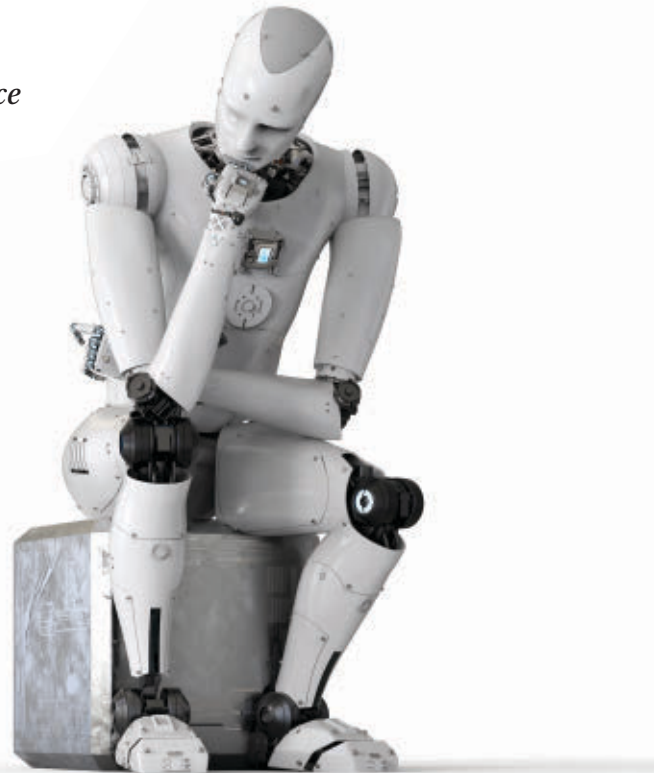


IMAGE BY PHONLAI PHOTO



with issues of representation and how we reason in language,” said Ellie Pavlick of Brown University in Providence, RI. “Does succeeding on [a specific] test mean you’re succeeding at language, does it mean you’re succeeding at common sense, or somewhere weird in between?”

One revealing example is the “Winograd Schema Challenge,” proposed in 2012 as an improvement on the venerable Turing test. This task requires a system to resolve what a pronoun refers to, like “they” in the sentence “The city councilmen refused the demonstrators a permit because they {feared, advocated} violence.” The sensible answer depends on whether the verb is “feared” (councilmen) or “advocated” (demonstrators). Nothing in the syntax dictates the choice. “We believed that in order to solve the Winograd Schema challenge, a system would need commonsense reasoning,” said Leora Morgenstern, now at PARC in Palo Alto, CA. “This is not the case.”

In constructing such ambiguous sentences, Morgenstern and others seek to avoid answers in which the alternative word choices would be statistically correlated with the correct pronoun assignment. Such statistical analysis, trained on enormous databases, is used in large-scale machine-learning language models like BERT, developed by Google, which are extremely good at exploiting correlations in the occurrence and arrangement of words in written language.

Contrary to their initial expectation, however, “When you look at this vast trove of sentences that have been collected, the statistical information has been captured” that is needed for disambiguating pronouns, Morgenstern said. “It reflects, in a way, some part of the commonsense knowledge that exists.” Still, she says the challenge fell short because these systems clearly do not have common sense, which would connect words or concepts that are widely separated in the text, or perhaps not even written down.

Although researchers have developed lots of “challenge” databases to try to measure common sense, “I don’t think we have very good metrics,” said Ernest Davis of New York University. “That’s part of the problem with this research area.”

Developing assessments is an important goal, agreed Matt Turek, who runs the Machine Common Sense program

**“We believed that in order to solve the Winograd Schema challenge, a system would need commonsense reasoning. This is not the case.”**

at the U.S. Defense Advance Research Projects Agency (DARPA). Just a year or so into the four-year program, some participants had already achieved most of the program goals for existing benchmarks for common sense. “What’s really driven that level of results is the rapid increase in the capability of these large-scale machine-learning-based language models,” he said. “That doesn’t mean that they’ve learned common sense.”

Turek said multiple-choice or true/false questions are not as informative as tasks that require generating new answers. It is hard to score such unstructured responses at large scale, however, which is critical for providing feedback for machine learning.

### Competing Representations

The longest-standing approach to embodying common sense does not depend on automated training, but on explicit, symbolic rules. These relationships often are represented as knowledge graphs in which nodes describing a concept are connected by arrows describing their relationship, such as “Napping” → “Causes” → “Energy.”

The CYC system, which Lenat has been working on for decades, extends this tradition of formal representation. He stresses, however, that binary relationships, and even third- and fourth-order relationships, are not rich enough. For example, in a *Communications* article, “You have no trouble following five, six, seven-deep nested modals about expectation, belief, desire, opposition to, and so on,” Lenat said. “Those are exactly the kinds of things that, kicking and screaming, we were led to represent in building the CYC system in the 1980s,” using formal

## ACM Member News

**AT THE INTERSECTION OF VISUALIZATION, COMPUTER GRAPHICS, AND COMPUTER VISION**



“My high school had a Commodore PET, and that was my first introduction to computing,”

recalls Hanspeter Pfister, An Wang Professor of Computer Science in the School of Engineering and Applied Sciences at Harvard University.

That introduction had Pfister wanting to learn how to build computers, so he decided to study electrical engineering. Pfister earned both his undergraduate and master’s degrees in electrical engineering at ETH Zurich, Switzerland, then went to Stony Brook University in New York for a Ph.D. in computer science.

On graduating in 1996, he joined Mitsubishi Electric Research Laboratories, where he rose to the roles of Associate Director and Senior Research Scientist. Pfister was chief architect of VolumePro, Mitsubishi Electric’s real-time volume rendering hardware for PCs, which he considers one of his top professional achievements.

His research in visual computing lies at the intersection of visualization, computer graphics, and computer vision. It spans topics including biomedical analysis and visualization, image and video analysis, and visual analytics in data science.

“Biomedical image analysis is where I have a lot of passion and can make an impact,” Pfister says. His current focus is on cancer cell analysis, the analysis of embryo cells for in vitro fertilization, and the analysis of neurons and their synaptic connectivity from electron microscopy data.

Pfister feels interpretable machine learning and transparency for deep learning models are important in order for doctors and clinicians to trust deep learning methods. “The first topic drives the second, and they go hand in hand.”

—John Delaney

representations of higher-order and modal logic. “The rest of the world is sort of stuck back where Marvin Minsky was in 1965,” he said.

Drawing inferences from these complex relationships is computationally challenging, however. “Over the last 35 years, we’ve identified about 1,000 different ways of speeding up logical inference so that our systems can do complicated reasoning in real time,” Lenat said. His company boasts clients in specialized applications like military and medical uses.

Because these projects are proprietary, however, other researchers cannot easily assess them, said Yejin Choi of the University of Washington and the Allen Institute for Artificial Intelligence. Choi noted that representing knowledge in such abstract constructs makes it hard for others to interpret or augment it, a possibility she exploits by crowdsourcing her work using Amazon Turk. “Natural language is far more expressive than what we know how to describe only using logical forms,” she said.

In contrast, Pavlick notes that “you can have common sense without any ability to speak language.” Despite her background studying natural language, she has begun exploring “reference-heavy” systems that use virtual worlds to develop “the notion of the things that language refers to and then learn language to refer to those things.”

### Prospects

Despite their differing preferred representations, many researchers agree that success in machine common sense will depend on combining different approaches. In the DARPA program, Turek said, “Some of the interesting work is on the interplay between graph representations—which have been around for a while and are quite rich, and allow you to do various types of reasoning—and deep learning, which might give you a good feature representation for text or images.”

Similarly, the human brain’s two hemispheres combine intuitive heuristics with formal reasoning, Lenat said. “Most AI systems in the future, even in the near future, will have the same kind of architecture.”

A recent project from Choi’s team called COMET (for Commonsense Transformers for Automatic Knowl-

**“One of the things that’s really inspiring about human infants is their ability to get broad general knowledge and then apply that successfully to specific challenges.”**

edge Graph Construction) incorporates a semi-structured knowledge graph with self-supervised machine learning, like that underlying large-scale language models. COMET has earned praise for plausibly completing sentences, which Choi said is a better task than teaching AI systems “to cheat better on multiple-choice questions.”

Elemental Cognition also is trying “to build a machine that does both” statistical language generation and capturing a “fundamental reality that causes the language,” Ferrucci said. In contrast to Watson, “I ultimately want to produce a model that is aligned with how humans acquire, structure, and communicate information.” Part of that process is using interactive exchanges with human users to “develop and maintain a shared understanding of the world,” he said.

Such interactive learning, inspired by the way children learn, is also one of two thrusts of the DARPA program, Turek noted. “There are really fundamental questions at play, both for the child developmental psychology community and how we might apply to that to inspire a new generation of artificial intelligence techniques,” he noted. “One of the things that’s really inspiring about human infants is their ability to get broad general knowledge and then apply that successfully to specific challenges.”

“Humans are often very good, if they have the right kind of framework, at learning from a single example,” Davis agreed, unlike deep-learning systems, which demand enormous

datasets. A child exposed to an iguana, for example, can immediately deduce that, like other animals, an iguana is born small and will eventually die.

Pavlick notes that better training algorithms could help machine-learning systems generalize. “Their training isn’t really set up to incentivize them to learn the kind of representations ... that would allow them to do this kind of quick generalization.”

She warns, however, that training can easily encode prejudice. “It’s hard to come up with a clear way of differentiating between the OK kinds of probabilistic associations and inferences that people make in common sense and ones that are just bad stereotypes.”

One advantage of including a formal component is that it inherently produces a rigorous explanation for its conclusions, Lenat said, not just a post-hoc rationalization. “It’s been understood for a very long time that common sense involves both the knowledge base and the inference ability,” agreed Morgenstern. “There has been a shift in emphasis,” she said, but “you need both.” **Q**

### Further Reading

Davis, E. and Marcus, G.  
Commonsense reasoning and commonsense knowledge in artificial intelligence, *Communications of the ACM* 58, No. 9, August 2015, <https://dl.acm.org/doi/10.1145/2701413>

Davis, E. and Marcus, G.  
Rebooting AI: Building Artificial Intelligence We Can Trust, Pantheon Books, New York, 2019, <https://amzn.to/37YBkke>

Kocijan, V., Lukaszewicz, T., Davis, E., Marcus, G., and Leora Morgenstern, L.,  
A Review of Winograd Schema Challenge Datasets and Approaches, April 23, 2020, <https://arxiv.org/abs/2004.13831>

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y.  
COMET: Commonsense Transformers for Automatic Knowledge Graph Construction, <https://arxiv.org/abs/1906.05317>

### Suggested Videos

CACM Sept. 2015 - Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence  
David Ferrucci, Machines As Thought Partners

Don Monroe is a science and technology writer based in Boston, MA, USA.

© 2020 ACM 0001-0782/20/11 \$15.00

# Natural Language Misunderstanding

*How do we eliminate bias in automated speech recognition?*

**I**N TODAY'S WORLD, it is nearly impossible to avoid voice-controlled digital assistants. From the interactive intelligent agents used by corporations, government agencies, and even personal devices, automated speech recognition (ASR) systems, combined with machine learning (ML) technology, increasingly are being used as an input modality that allows humans to interact with machines, ostensibly via the most common and simplest way possible: by speaking in a natural, conversational voice.

Yet as a study published in May 2020 by researchers from Stanford University indicated, the accuracy level of ASR systems from Google, Facebook, Microsoft, and others vary widely depending on the speaker's race. While this study only focused on the differing accuracy levels for a small sample of African American and white speakers, it points to a larger concern about ASR accuracy and phonological awareness, including the ability to discern and understand accents, tonalities, rhythmic variations, and speech patterns that may differ from the voices used to initially train voice-activated chatbots, virtual assistants, and other voice-enabled systems.

The Stanford study, which was published in the journal *Proceedings of the National Academy of Sciences*, measured the error rates of ASR technology from Amazon, Apple, Google, IBM, and Microsoft, by comparing the system's performance in understanding identical phrases (taken from pre-recorded interviews across two datasets) spoken by 73 black and 42 white speakers, then comparing the average word error rate (WER) for black and white speakers.

The subjects used in the recordings found in the first dataset were from Princeville, a predominantly African-American rural community in North Carolina; Rochester, a mid-sized city in western New York state, and the Dis-



trict of Columbia. The second dataset was the Voices of California, an ongoing compilation of interviews recorded across that state, although the focus was on Sacramento, the capital of California, and Humboldt County, a predominantly white rural community in northern California.

The researchers indicated that black subjects spoke in what linguists refer to as African-American Vernacular English, a variety of English sometimes spoken by African-Americans in urban areas and other parts of the U.S. This is contrasted with the Standard English phrasing most often used by white speakers.

Overall, the researchers found the systems make far fewer errors with users who are white than with users who are black. ASR systems misidentified words about 19% of the time with white speakers, with the WER rising to 35% among black speakers. Approximately 2% of audio snippets from white people were considered unreadable by these systems, compared with 20% of snippets spoken by black people.

“Our paper posits that much of the disparity is likely due to the lack of training data on African Americans

and African American Vernacular English speech,” explains Allison Koencke, a Stanford doctoral student in Computational Mathematics & Engineering, and the first author of the study. “It seems like the lack of training data is in particular traced to disparities arising from the acoustic model, as opposed to the language model.”

Acoustical training models are focused on correctly understanding words despite differences in accents, speech patterns, tone of voice, and diction, compared with language models, which are designed to recognize various words and phrases used by speakers. According to the study, “Our findings indicate that the racial disparities we see arise primarily from a performance gap in the acoustic models, suggesting that the systems are confused by the phonological, phonetic, or prosodic characteristics of African American Vernacular English rather than the grammatical or lexical characteristics. The likely cause of this shortcoming is insufficient audio data from black speakers when training the models.”

The key to improving ASR accuracy among all speakers is to use a more diverse set of training data, which should include speakers that come from more diverse ethnic, cultural, and regional backgrounds, according to Sharad Goel, a co-author of the study and an assistant professor of management science and engineering at Stanford.

“We have tried to stay away from the blame game and say, ‘oh, we think you’re like, you know, good or bad because you didn’t prioritize it,’ but we really think this is important,” Goel says. “We hope people will change their behavior, especially these five companies, but also more broadly in the speech recognition community, toward improving these outcomes.”

ASR technology companies may be hearing that message loud and clear. An Amazon spokesperson pointed to a statement published after the release of the Stanford study, which noted that “fairness is one of our core AI principles, and we’re committed to making progress in this area ... In the last year we’ve developed tools and datasets to help identify and carve out bias from ML



models, and we offer these as open source for the larger community.”

Other vendors that utilize ASR technology say that despite their complexity and capabilities, ML models require a good deal of human oversight, particularly as models are trained. In some cases, ASR technology developers would use a relatively limited range of voices, speech patterns, or accents to train their acoustical models, with the goal of rapidly developing a solution that could be commercially deployed. While this approach may yield a high degree of accuracy with neutral speakers, it may struggle with accents or dialects that differ from the voices used to train the model.

“So, you could build out a quick and dirty solution that is very powerful, but it would fold over at the first hurdle because it doesn’t understand the accent, doesn’t understand the terminology, doesn’t even understand my language, and so on and so on,” says Andy Peart, chief marketing and strategy officer at Artificial Solutions, Stockholm, Sweden-based developer of the Teneo enterprise-focused conversational platform. “We would argue that you need to think about all these things to build out something that’s actually going to be effective.”

Peart says Artificial Solutions uses a hybrid ML approach to training. ML is used for the initial training of the models, but human engineers are deployed to make sure that the system continually learns on the right inputs, which can include matching speaker voice inflections and pronunciations to the appropriate words or intents.

Further, the system is designed to assign a confidence ratio to the accuracy of the ASR model as applied to voice inputs. If the confidence ratio is below a certain threshold, the system is designed to ask the speaker for clarification, such as by asking, “did you mean \_\_\_\_\_?”

“We don’t settle for learning [solely] within the solution, because then you potentially get the Microsoft Tay situation, where your solution automatically learns and changes from the inputs without any control from the company. This would be catastrophic in a commercial environment,” Peart says, referring to the ability of users to train the Tay unsupervised ML-based chatbot to

spew racist and otherwise offensive content, based on voice and text inputs and a lack of moderation of the machine’s responses by human engineers.

Other ASR vendors note the initial training data should be diverse, in order to function accurately for all types of users. “In order to train really good machine learning models, you need a large amount of data, but you also need diverse data,” says Johann Hauswald, co-founder and chief customer officer with Cline, a conversational AI platform provider based in Ann Arbor, MI.

“We recommend customers use crowdsourcing platforms to collect training data,” Hauswald says, citing as examples Amazon Mechanical Turk and CrowdFlower (now Figure Eight), which include more diverse speaker data. “We take the approach of crowdsourcing that [training] data and not [relying solely on] a small set of folks collecting and training our data.”


Hauswald says the other advantage of using data from crowdsourced platforms is the ability to collect a wider range of words or phrases that mean the same thing, thereby expanding the lexicon of the ASR system (such as correctly identifying that “y’all” is a shortened, slang version of “you all” in Southern U.S. dialects). He notes the platforms ask the same question across a broad, diverse range of speakers, which increases the depth of the training model to account for ethnic, regional, gender, and other differentiators.

“You get a large amount of data, but then also from a diverse set of people,” Hauswald says, “It’s not one person giving you 500 utterances, and it’s not 500 people giving you a single [phrase].”

According to Hauswald, ASR systems struggle with heavily accented speech simply because there is significantly more training data consisting of non-accented English than there is for foreign or minority accented languages. Hauswald says ASR algorithms identify speech by looking for sound patterns, then linking them to appropriate words, which requires some human intervention in order to ensure that even when sounds are mispronounced (such as ‘r’ sounds being pronounced as ‘l’ sounds), the correct word is chosen. With less available foreign-accented data to analyze, it becomes more difficult to identify

patterns that can be used to train the model accurately. One solution is to simply collect and train ASR models using speech data from accented speakers, and then using humans to ensure that the model correlates the accented pronunciations with the correct words. However, collecting enough speech data from each type of individual accent is fraught with compute, time, and data-collection challenges.

One way to speed up this process is to utilize a concept called transfer learning, a technique in which an ASR model is trained on a large set of data, such as speakers using un-accented English. The basic techniques the model uses to learn specific phonetic and speech data patterns then can be applied to a second, smaller dataset containing accented-English speech. The parameters and techniques from the first dataset are used as a starting point for training on the second dataset, which speeds up the learning process, allowing the new model training to focus on the unique pronunciations found in the accented speech.

“For languages or dialects that have less training data, research has shown you can use a language that has more data and use transfer learning to refine a model for the target language,” Hauswald says. He explains that approach has become popular, “initially in image processing, but now the same techniques are being applied to natural language processing and speech recognition pretty successfully. But you still need to go through that step of kind of hand annotating, sanitizing, and cleaning that data.” 

#### Further Reading

Racial Disparities in automated speech recognition, *Proceedings of the National Academy of Sciences of the United States of America*, April 7, 2020. <https://doi.org/10.1073/pnas.1915768117>

Gender and Dialect Bias in YouTube’s Automatic Captions, *Conference Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, January 2017. DOI: 10.18653/v1/W17-1606

Acoustic Modeling Explained: <https://www.youtube.com/watch?v=5ktDTa8glaA>

Keith Kirkpatrick is principal of 4K Research & Consulting, LLC, based in Lynbrook, NY, USA.

# Terahertz Networks Move Closer to Reality

*The desire for faster, higher-frequency wireless networking is a constant. Terahertz technology could deliver large gains.*

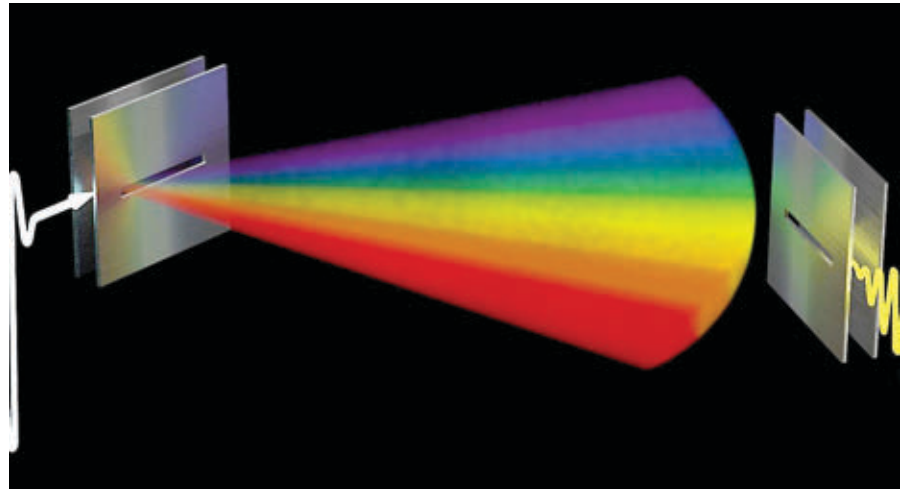
**W**IRELESS TECHNOLOGIES HAVE advanced by leaps and bounds over the last several decades. Speeds have increased dramatically, connectivity has improved, and wireless network protocols—including Wi-Fi and cellular—have become ubiquitous. Today, a reasonably fast, persistent wireless connection is available in most parts of the world.

Yet, despite all the gains, today's wireless networks are still relatively limited in terms of how they handle large volumes of data. Connection speeds are adequate for basic tasks like Web browsing, videoconferencing, e-commerce, gaming, streaming videos, and exchanging messages. However, as new and more resource-intensive technology enters the picture and the Internet of Things (IoT) expands, the need for bandwidth is growing—in some cases, by an order of magnitude.

Terahertz communications, also known as submillimeter radiation, could address this problem. Because it relies on the high end of the radio spectrum, at frequencies from 300GHz to 1,000GHz, it delivers far more bandwidth than today's networks, and can connect many more devices. At the same time, terahertz technology has characteristics that could radically change the nature of computing and the IoT. This includes sophisticated sensing capabilities based on spectroscopic signatures.

Recently, terahertz communication took a major step toward becoming a viable technology. Researchers at Brown and Rice universities found a way to solve a longtime problem with keeping devices connected to a terahertz network. As a result, optimism is growing that the technology will soon evolve to a commercial scale.

"It's possible to do a lot more things—and introduce more advanced



**Terahertz frequencies escape from leaky waveguides in a 'rainbow' pattern that can be used to help devices find and identify each other in future terahertz wireless networks.**

features—at higher frequencies," says Daniel Mittleman, a professor of engineering at Brown University. "Terahertz could be transformative."

## Crossing the Spectrum

The appeal of terahertz networking is clear: today's cellular and Wi-Fi networks carry voice conversations and data streams using microwaves. However, the lower end of the radio spectrum cannot accommodate the growing demand for bandwidth. As networks become overwhelmed, speed and device performance degrade. Even emerging 5G (fifth-generation cellular technology) has limitations and eventually will hit technical and practical limits for speed and performance. On the other hand, terahertz waves, which exist between optical and microwave frequencies, increase transmission speeds multifold, while expanding what wireless systems can do.

Terahertz networking "represents the next leap forward of what we can achieve with communication and sensing," says Edward Knightly, Sheafor-Lindsay Professor of Electrical and Computer

Engineering at Rice University.

Future routers and 6G cellular networks likely will incorporate terahertz spectrum. However, that is just the start. By providing access to almost unlimited bandwidth, terahertz systems could allow fleets of autonomous vehicles or robotic devices, including drones, to communicate without bandwidth limitations that often lead to latency and network congestion; they could support massive virtual reality environments, and could revolutionize home health, physical security, and cybersecurity by introducing far more accurate motion and sensing capabilities. For example, the unique sensing capabilities of terahertz technologies could potentially monitor air quality and sense certain types of pollutants that may otherwise be difficult to identify, or scan people and bags at airports with far greater precision than today's millimeter-wave scanning systems.

In fact, sensing capabilities within networks could take a huge leap forward as a result of terahertz technology. "There are few, if any, materials that can be readily sensed using spectros-

copy in the lower-frequency bands that are traditionally used for radar or wireless communications,” Mittleman says. “But the terahertz range is rich with spectroscopic signatures, so there are many feasible sensing targets.”

It’s an important consideration. “The technology creates completely new opportunities for detecting motion and for using indoor positioning,” says Yasaman Ghasempour, an assistant professor of electrical engineering at Princeton University who studied terahertz technology as a graduate student at Rice University. For example, terahertz sensors could detect when someone or something that should not be there enters a specific space or area, with high degree of accuracy. In addition, unlike vision-based techniques, terahertz communication is unaffected by lighting or atmospheric conditions, and it can better preserve privacy, she explains. For instance, it’s possible to tune terahertz transmissions to a specific distance, say 100 or 200 meters, and thus control who has access to the signals, Mittleman says.

Still another aspect of terahertz radiation that’s particularly intriguing is its ability to bounce signals off reflective surfaces. Ghasempour says certain materials such as metals, glass, or even concrete can create strong reflections

## Terahertz signals bounce off reflective surfaces; some materials can create reflections that can extend gigahertz coverage across houses, buildings, and factories.

that extend terahertz coverage across houses, buildings, and factories. As terahertz technology is incorporated into cellular towers and routers, “This would make it possible to reduce or eliminate antennas, range extenders, access points, mesh systems, and other assistive wireless technologies. The reflective surfaces serve as an antenna instead,” she points out.

### All About Connections

Terahertz sensing already is used for specialized applications, such as satellites monitoring the atmosphere,

and specialized short-range imaging and signal transmission. Yet, the physics of the terahertz spectrum has built-in limitations for networking. For one thing, terahertz wavelengths propagate in a completely different way than microwave technology, and they cannot penetrate solid walls or pass through many objects, such as steel and wood (they are able to pass through some types of glass and plastic). Unlike current wireless technologies such as 802.11 or cellular networks, the waves cannot saturate an area or indoor space and easily connect to devices. “The question becomes what is the right direction to point the signal?” says Josep Jornet, an associate professor in the department of electrical and computer engineering at Northeastern University.

At frequencies in the millimeter-wave range, conventional device discovery takes place through a sequential search process. A transmitting device, say a router or cell tower, scans at various angles until it finds a device with which it can connect. However, as the frequency increases the beams become narrower, and more steps are required for this discovery process. This renders the sequential approach impractical. The problem is magnified when large numbers of mobile de-

# ICCQ

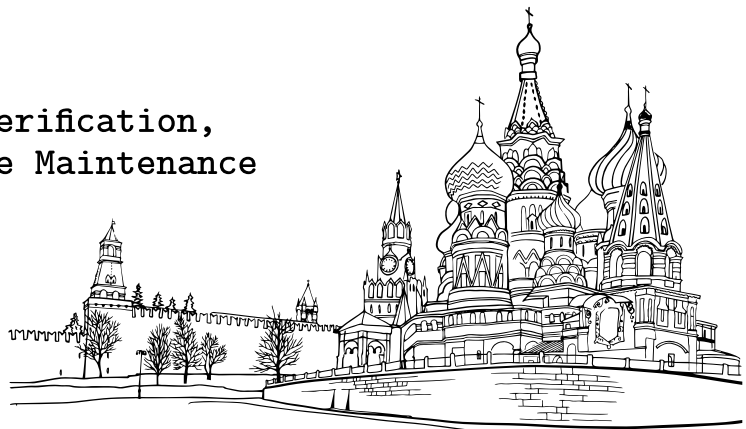
[www.iccq.ru](http://www.iccq.ru)

The First International Conference on Code Quality  
in cooperation with IEEE Computer Society

Moscow, Russia  
27 Mar 2021

Static Analysis, Program Verification,  
Bug Detection, and Software Maintenance

CfP closes:  
4 Dec 2020





vices become part of a network. “Finding the right direction to transmit the signal between the sender and receiver can be difficult and time-consuming. If all clients are moving, the potential overhead is high just to keep the beams aligned and devices connected,” Knightly points out.

Researchers are taking direct aim at these technical barriers. For instance, the group at Brown and Rice universities demonstrated that a device called a *leaky waveguide*, which has been used in other microwave systems, can assist with link discovery at terahertz frequencies. The passive device is comprised of two metal plates that allow radiation to propagate in the space between them. One of the plates has a small opening that allows small amounts of radiation to leak out. By changing the opening, it is possible to adjust the frequency of the input signal to the waveguide. As a result, radiation escapes in different directions. As a broadband signal enters the device from a router or cell tower, the device creates a unique signature. Once this handshake takes place, the router or cell tower can infer directional information, lock onto it for directional communication, and maintain a persistent connection as the user and device move about.

Reflective materials also could solve part of the transmission problem and boost router and network sensing capabilities. “By adding special substances to walls, floors, and ceilings in rooms and using various objects, such as lamps and wall paintings, to bounce signals throughout the structure, you essentially potentially wind up with hundreds or thousands of antennas and potential sensors,” Ghasempour says. This capability, along with improved algorithms designed to interpret data and reflective angles, could produce a faster and much smarter communications network than what’s possible today. “There is a lot of additional research that needs to be done on reflective materials, but they are clearly a viable tool for developing terahertz technology,” she says.

Yet another area of current research focuses on chips and antennas. At Nanyang Technological University in Singapore, researcher Ranjan Singh is studying the use of topological photonics to address in-

## Key challenges remaining include how to maintain persistent connections to large numbers of devices on the network, and clearly defining industry standards.

herent problems with intrachip and interchip communication at terahertz frequencies. This includes creating new designs and shapes that better accommodate the signals. Jornet, who has studied terahertz technology for more than a decade, has examined everything from materials and antenna designs to new algorithms and routing methods. “The biggest remaining question is how to produce a compact chip, radio, and communications protocol that can work in the real world within an integrated platform. Ultimately, terahertz networking is both a physics problem and a communications problem.”

### A New Wavelength

Commercial products using terahertz technology will likely appear sometime over the next five to 10 years, Knightly says. Solving the leaky waveguide problem was a major step forward, but there are still several key challenges to overcome. These include issues ranging from how to maintain persistent connections when large numbers of devices are connected to terahertz networks, to more clearly defining industry standards. There’s also a need to add outdoor infrastructure, including the smaller, more densely packed cells that terahertz networks require. Nevertheless, “Terahertz’s unique capabilities of sensing and high-rate low-latency networking will drive commercialization,” Knightly says.

In fact, Knightly and others believe terahertz communications ultimately will be a game-changer. It will deliver the ultra-high bandwidth and, as a result, the super-low latency needed to operate autonomous machines and other systems used in smart factories and cities, through Wi-Fi and 6G. It will add sophisticated sensing features that may reshape networking, and it could transform virtual reality from a niche technology into a mainstream tool that complements or replaces today’s teleconference and presence systems.

Says Knightly, “Terahertz supports large numbers of people and environments that completely overwhelm today’s technology. This opens up all kinds of possibilities.” □

### Further Reading

Ghasempour, Y., Shrestha, R., Charous, A., Knightly, E., and Mittleman, D.M., **Single-shot link discovery for terahertz wireless networks**, *Nature Communications*, April 24, 2020, Vol. 11, 2017. <https://doi.org/10.1038/s41467-020-15761-4>

Amarasinghe, Y., Mendis, R., and Mittleman, D.M. **Real-time object tracking using a leaky THz waveguide**, *Optics Express*, 2020, Vol. 28, Issue 12, pp. 17997-18005. <https://www.osapublishing.org/oe/abstract.cfm?uri=oe-28-12-17997>

Xia, Q., Hossain, Z., Medley M. J., and Jornet, J.M., **A Link-layer Synchronization and Medium Access Control Protocol for Terahertz-band Communication Networks**, *IEEE Explore*, September 10, 2019. <https://ieeexplore.ieee.org/abstract/document/8827939/authors#authors>

Karl, N.J., McKinney, R.W., Monnai, Y., Mendis, R., and Mittleman, D.M., **Frequency-division multiplexing in the terahertz range using a leaky-wave antenna**, *Nature Photonics*, September 14, 2015, Vol. 9, Pages 717-720. <https://www.nature.com/articles/nphoton.2015.176>

Xia, Q., Hossain, Z., Medley M. J., and Jornet, J.M., **Expedited Neighbor Discovery in Directional Terahertz Communication Networks Enhanced by Antenna Side-Lobe Information**, *IEEE Transactions on Vehicular Technology*, Aug. 2019, Vol. 68, Issue: 8. <https://ieeexplore.ieee.org/document/8746670>

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

© 2020 ACM 0001-0782/20/11 \$15.00

# Privacy

## Digital Contact Tracing May Protect Privacy, But It Is Unlikely to Stop the Pandemic

*Considering the potential benefits versus the risks of privacy-enhancing technologies.*

**I**T IS DIFFICULT to imagine a timelier topic for this inaugural *Communications Privacy* column than the privacy issues associated with COVID-19 apps. Against the backdrop of protests around the world opposing racism and police killings of Black people, we have a newly found understanding of the need for protection from surveillance, while also feeling the urgency of shutting down the spread of a deadly virus. While many computer scientists are looking to technology for privacy-protective ways to track COVID-19 exposure, Privacy-enhancing technologies (PETs) may prove ineffective without more widely available COVID-19 tests, human-centered design, and complementary laws and policies.

As the COVID-19 pandemic spread in spring 2020, researchers and public health officials pursued digital contact tracing and exposure notification tools to assist human contact tracers. Initial efforts to build these tools focused on

utility but were quickly met with questions about privacy. Although there is compelling public interest in sharing data to reduce virus spread, concerns arose that this data might be used for other purposes. Indeed, as protest marches became commonplace and police sought out instigators, rumors spread that police might be using data collected by contact-tracing apps. While I have seen no evidence that this actually occurred, the concern is legitimate and may slow app adoption. In the U.S., people of color have been disproportionately hard hit by COVID-19 but may also have the most to fear in using these apps.

Digital contact tracing and exposure notification might be ideal applications for PETs. An app that could notify users that they had been exposed to COVID-19 without leaking locations or personal information could simultaneously protect both public health and privacy. However, PETs alone may not solve this problem.

### Digital Contact-Tracing Technology

Contact tracing and exposure notification apps run on mobile phones and log either the phone's location or the presence of other nearby phones. The location approach involves sending location information to a centralized server so that users who were recently co-located with a user who tested positive for COVID-19 can be identified. Alternatively, an infected user's locations could be broadcast to other users so that their apps can check for co-location. The proximity approach does not require storing location and instead logs an identifier associated with each phone the app detects as being nearby for some period of time (for example, 10 minutes). If a user tests positive, their identifier is broadcast, and apps can check to see whether the infected person's identifier is in their proximity logs. The proximity approach is more privacy protective as it determines only that two users were near each other, not all the places they were located.



Research teams around the world have been working on privacy protective protocols for contact tracing and exposure notification. These include the Private Automated Contact Tracing (PACT)<sup>a</sup> group led by Massachusetts Institute of Technology researchers and the European Decentralized Privacy-Preserving Proximity Tracing (DP-3T) consortium.<sup>b</sup>

Google and Apple jointly developed a Bluetooth-based “Exposure Notification” API for Android and iOS platforms that public health agencies can incorporate into contact-tracing apps. It uses a decentralized and privacy-protective approach, which includes cryptographically generated rotating identifiers that make it more difficult (but not impossible<sup>1</sup>) to trace an identifier back to an individual.

Exposure Notification has not yet been built into many apps, although new apps that use this API are expected to launch throughout fall 2020. Many apps seem to be using their own approaches,

a See <https://pact.mit.edu/>

b See <https://github.com/DP-3T/>

and privacy and security issues are common. For example, concerns have been raised about apps that may leak sensitive information through security holes. A mid-June report<sup>3</sup> assessed mobile-phone-based contact-tracing apps from government entities around the world and found most were susceptible to being tampered with to allow attackers access to sensitive data.

### Trade-Offs

Some apps use good technical approaches to limit data leakage, but in doing so, they may limit their utility. I can imagine an app informing me that sometime in the past week, or more precisely last Tuesday, I was in proximity of someone who has tested positive for COVID-19. I would have questions. How long was I near them? Were we outdoors? Were they coughing? Were either of us wearing a mask? Were we talking face-to-face or standing silently six feet apart? Or were they on the other side of a solid wall? I could imagine feeling upset and wanting to talk to a human rather than receive a notification from an app.

I would likely pay attention to my first positive notification from a contact-tracing app. I might seek out a COVID-19 test and would probably quarantine myself. But should I stay away from the other members of my household? Should they quarantine themselves too? (More questions I might want to ask a human.) After the first notification, I might have more experience and know what to do, but I would probably soon start to ignore these notifications, assuming them to be false positives. This problem is exacerbated by the lack of widely available rapid COVID-19 tests in most of the U.S. and many other countries.

There are ways to design an app to answer many of my questions, mitigating some concerns. This may require that we give up some privacy. Maybe the app should store both location information and proximity information so it can communicate where the exposure occurred. Maybe the infected person could grant permission for additional information—such as whether they wore a mask in public—to be transmitted to people receiving notifications. Maybe they would consent to sharing their name with friends who



were receiving notifications. Laws and policies restricting the use of contact-tracing information might reduce privacy concerns and encourage people to allow more data to be collected and shared. In addition, people may be willing to allow an app to collect more data in certain places—I might allow an app to collect precise location information at the grocery store or park, but not at a doctor's office or protest march. However, it may be challenging to build an app that supports this sort of control without requiring users to spend a lot of time and effort setting it up.

Another concern is people may falsely report they have been infected to cause mischief or to keep people home in order to shut down school or even to disrupt an election. This problem is being addressed by requiring public health officials, doctors, or testing labs to verify positive test reports before notifications are sent, although this may reduce convenience, privacy, or timely notification.

### Adoption

A study conducted last spring in the U.S. found participants generally preferred a centralized contact-tracing approach that would share identities and location of infected users with public health authorities rather than a decentralized and more privacy-protective approach that did not share data. Approximately half of the participants reported a willingness to install such a centralized app, while about a quarter of the participants indicated they would be unlikely to install any contact-tracing app. Other U.S. surveys have also found that about half of smartphone users report being likely to install a contact-tracing app.<sup>6</sup> This does not take into account the people who do not own smartphones and thus would not be able to use these apps. In the U.S., some of the demographic groups most at risk are least likely to own smartphones. Furthermore, current adoption of these apps appears low, even in countries that introduced apps last spring. For example, in May it was reported that contact-tracing apps had been adopted by only 38% of the population in Iceland and 20% in Singapore. Yet researchers estimate adoption rates of at least 60% are necessary for these apps to be effective.<sup>5</sup> Even in France where the StopCovid app was activated by over 1.8 million people in June, the app notified only 14 people that they may have been exposed.<sup>1</sup>

At Carnegie Mellon University, a team of researchers developed an anonymous contact-tracing app called NOVID that distinguishes itself from other apps by using ultrasound in addition to Bluetooth to improve accuracy and allow it to provide notification recipients with information about how close they were to an infected person. One of its features is that it tells users if they are in proximity of other NOVID users even when no infection has been reported. I live in a neighborhood near Carnegie Mellon, where there is likely more interest in NOVID than in other places. In the four months since I installed NOVID, it has detected that I have been near only one other NOVID user, despite the fact I have brief contact with numerous people through daily outdoor exercise and regular errands. Thus, NOVID is not yet particularly useful to me.

While public adoption has been slow, private companies and universities are starting to mandate their employees and students use apps to trace contacts and report symptoms. Some companies are mandating the use of apps or wearable devices that immediately alert wearers when they are too close to other users. Symptom-tracking apps can provide a short daily questionnaire for users to report any COVID-related symptoms. However, all of these approaches are raising concerns. Employees and students wonder where their information is sent and how it can be used. The University of Connecticut conducted focus groups and found students were unlikely to report symptoms such as headaches that occur frequently for reasons unrelated to COVID-19, for fear of being forced to quarantine and miss exams or social events.<sup>c</sup>

In contrast to apps, wastewater monitoring may be more privacy-protective (assuming samples are taken as water exits the building rather than with every flush), easier to deploy, and more effective at detecting COVID at universities, providing an early warning when someone in a monitored building is infected. All the occupants of a University of Arizona dorm were tested after the virus was found in their building's wastewater in August. As a result two asymptomatic students were discovered before they spread the virus further.<sup>7</sup>

c See <https://bit.ly/2FAEJLM>

### Challenges

While efforts to use apps to help control a deadly virus and protect privacy are laudable, early efforts do not look promising, and some experts have concluded the risks of contact-tracing apps might outweigh their potential benefits.<sup>2,8</sup> Challenges remain in developing and deploying widely apps that are highly effective and usable.

Technologists have focused on building privacy-protective decentralized apps, but a centralized approach with legal protections to limit data use might be more beneficial to public health and more understandable and acceptable to the public. The problem these apps are trying to solve is not just a technology problem, and digital technology alone is unlikely to be the solution. As with many privacy problems, solutions should involve both policy and technology.<sup>9</sup> Laws and organizational policies must ensure sensitive information is not used for purposes unrelated to public health; digital tools must be trustworthy, understandable, and usable; and public health organizations and rapid COVID-19 testing must be part of the solution. **C**

### References

1. Dillet, R. French contact-tracing app StopCovid has been activated 1.8 million times but only sent 14 notifications. *TechCrunch* (June 23, 2020); <https://torn.ch/3bW1g1P>
2. Gebhart, G. COVID-19 tracking technology will not save us. *Electronic Frontier Foundation* (Sept. 3, 2020); <https://bit.ly/3mo8jF9>
3. Goodes, G. Report: The Proliferation of COVID-19 Contact Tracing Apps Exposes Significant Security Risks. *Guardsquare*. June 18, 2020; <https://bit.ly/2E0bjX7>
4. Greenberg, A. Does Covid-19 contact tracing pose a privacy risk? Your questions, answered. *Wired* (Apr. 17, 2020).
5. Kreps, S. et al. Contact-tracing apps face serious obstacles. *Brookings Tech Stream*. (May 20, 2020); <https://brook.gs/2FCrGJE>
6. Li, J. et al. Decentralized is not risk-free: Understanding public perceptions of privacy-utility trade-offs in COVID-19 contact-tracing apps. (May 25, 2020); <https://bit.ly/2DXWF2q>
7. Peiser, J. The University of Arizona says it caught a dorm's covid-19 outbreak before it started. Its secret weapon: Poop. *The Washington Post* (Aug. 28, 2020).
8. Soltani, S., Calo, R., and Bergstrom, C. Contact-tracing apps are not a solution to the COVID-19 crisis. *Brookings Tech Stream*. (Apr. 27, 2020); <https://brook.gs/3iqvI6y>
9. Spiekermann, S. and Cranor, L.F. Engineering privacy. *IEEE Transactions on Software Engineering* 35, 1 (Jan.–Feb. 2009), 67–82; doi: 10.1109/TSE.2008.88.

**Lorrie Faith Cranor** ([lorrie@cmu.edu](mailto:lorrie@cmu.edu)) is Director and Bosch Distinguished Professor in Security and Privacy Technologies, CyLab Security and Privacy Institute and FORE Systems Professor, Computer Science and Engineering & Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.



## Legally Speaking

# Copyright's Online Service Providers Safe Harbors Under Siege

*Reviewing the most significant changes recommended in the recently released U.S. Copyright Office Section 512 Study.*

**F**OR THE LAST 20 years, the laws of the U.S., E.U., and most of the rest of the world, have provided Online Service Providers (OSPs) that host user-uploaded content with “safe harbors” insulating them from claims of copyright infringement so long as they did not know about or participate in infringing acts of their users. Yet, once a copyright owner has notified an OSP about the presence of infringing materials at a particular place on its site, the OSP has generally had a responsibility to investigate and remove or disable access to infringing materials. This notice-and-takedown safe harbor became part of U.S. law as a result of passage of the Digital Millennium Copyright Act (DMCA) of 1998.

This safe harbor, now codified in § 512 of U.S. copyright law, has never been popular with major copyright industries. Those industries acquiesced to these rules in 1998 as part of a grand compromise with other industries and organizations to get support for enactment of rules that outlaw tools for circumventing technical protection measures that major copyright industries planned to use when distributing digital copies of their works.

Copyright industry dissatisfaction with this safe harbor has grown increasingly strident because of the widespread prevalence of online infringe-



ments. As my November 2019 column reported, the E.U. responded to copyright industry complaints by adopting new strict liability rules for certain online content sharing services, which EU member states must implement by June 2021.

To aid possible Congressional reconsideration of the DMCA safe

harbors, the U.S. Copyright Office released in May 2020 its long-awaited Section 512 Study.<sup>2</sup> It sided with major copyright industry complaints on virtually every contentious safe harbor issue. This column reviews the most significant of the Study's recommended changes: tightened eligibility rules, revamped “red flag” knowledge of



infringement requirements, curtailment of fair use claims, and stricter repeat infringer policies.

Although the Study claims these changes would merely “fine-tune” § 512, this is a mischaracterization. Were Congress to follow the Study’s recommendations, OSPs would have much greater responsibilities to detect and thwart user infringements and more liability if they did not succeed in preventing copyright infringements. These changes would negatively impact the availability of user-generated content and other legitimate user activities. Stricter repeat infringer policies would likely mean that many more users’ accounts would be terminated, including those of innocent users.

### Eligibility for Safe Harbors

One very substantial change the 512 Study recommends would limit the availability of the hosting safe harbor to only those OSPs that passively store user contents. The Study criticized the *Viacom v. YouTube* decision for concluding that YouTube qualified for the § 512(c)’s safe harbor when proactively transcoding user-uploaded videos, enabling the videos to be viewed by others, and promoting user-uploaded videos through automated recommendations. The court regarded these activities as “related” to hosting user contents, and hence, within the safe harbor.

The CEO of the Internet Association, who testified before the Senate IP Subcommittee in June 2020, objected to the Study’s narrow interpretation of this safe harbor. Restricting its application would, he said, “all but exclude every modern OSP from the scope of section 512(c), giving liability protections only to the bulletin board services from the 1990s.”<sup>3</sup> Under this conception of § 512(c), YouTube, Facebook, Ravelry, Reddit, Hacker News, and every other social media service would lose the § 512(c) safe harbor completely.

He noted that “the DMCA was intended to incentivize innovation and the growth of the internet” and that “algorithmic recommendations—which benefit users by connecting them to their communities and information they are likely to be interested in—do not negate the principle that the underlying content is stored at the direction of the user.”<sup>3</sup>

Yet, the Study would strip OSPs of this shield from infringement claims, even when they neither knew about nor had encouraged infringing activities. OSPs would now be at risk of very large money damage awards instead of being subject only to the prospect of injunctive relief.

### “Red Flag” Knowledge and Willful Blindness

A second significant change recommended in the 512 Study concerns standards for judging when OSP hosting services should be deemed to “know” about user infringements. The Study claims that courts have misinterpreted what should constitute “red flag” knowledge and willful blindness to infringement.

Under the so-called “red flag” knowledge provision of § 512(c), OSPs lose the safe harbor when they are “aware of facts or circumstances from which infringing activity is apparent.” Once they have such awareness, they must act “expeditiously” to remove or disable access to infringing content. The *Viacom* decision said this rule applied only when OSPs became aware of specific and identifiable acts of infringement. (This ruling makes sense because OSPs cannot remove or disable access to infringing materials if they do not know what and where they are.)

The Study asserts that general knowledge of infringement somewhere on an OSP’s site should suffice as “red flag” knowledge. Even though § 512(m) provides that OSPs do not have a duty to monitor their sites to detect infringement, the Study regards this rule as protecting only user

**These changes would negatively impact the availability of user-generated content and other legitimate user activities.**

privacy interests. It argues that OSPs should be more proactive in monitoring their sites for infringements, even though § 512(m) says no such duty exists. Failure to monitor and investigate user-uploaded content if some OSP staff have some general awareness of infringement should result, the Study suggests, in OSPs being found to be willfully blind to infringement.

### What About Fair Use?

The 512 Study evinces a cavalier and largely dismissive attitude toward fair uses. It fails to recognize that widely posted user-generated content, such as remixes, mashups, and fan fiction stories, is generally viewed as making fair and non-infringing uses of copyrighted works when creatively reusing parts of popular music, videos, and the like.

Also generally fair are incidental uses, such as the short video that Stephanie Lenz uploaded to YouTube of her baby dancing with some Prince music playing in the background. Universal sent a takedown notice to YouTube about it, which Lenz counter-noticed on fairness grounds. An appellate court in *Lenz v. Universal Music Corp.* held that Universal must consider fair use in order to have a good faith belief that an online use of its work was infringement *before* sending a takedown notice. The 512 Study criticized *Lenz* as wrongly decided.

The Study also chided OSPs for deciding not to take down allegedly infringing content when persuaded that a challenged use was fair: “OSP’s do not appear to be fully honoring the requirement in § 512(c)(1)(C) that upon receiving a takedown notice that is compliant with § 512(c)(3), they ‘respond[] expeditiously to remove or disable access to’ the material.” Under the Office’s interpretation of § 512, OSPs must remove or block access to content about which a takedown notice has been received regardless of whether that use is fair.

### Repeat Infringer Policies

Under § 512(i), OSPs are eligible for § 512 safe harbors only if they have adopted and reasonably implemented a repeat infringer policy. The Study recognizes the need for some flexibility in the formulation of such policies and their implementations, in part because of the astonishing diversity of OSPs



these days. One size does not fit all.

Yet, the Study undercuts its endorsement of flexibility by arguing that all OSPs should have public formal written policies. It expressed disapproval of an appellate court ruling allowing a small provider to qualify because it reasonably honored takedown notices and terminated the accounts of users who repeatedly upload infringing files.

Another of the Study's troublesome conclusions is that accusations of infringement should suffice to deem users as repeat infringers. The Study dismissed as unrepresentative considerable evidence OSPs presented of DMCA takedown abuse. It mentioned, but did not heed, empirical evidence showing that nearly one-third of takedown notices are flawed because they were incomplete or fraudulent, the uploaded material was fair use, or the notice provider was not the owner of a copyright alleged to be infringed.

The Study contends that termination policies must apply to Internet access providers, as well as OSP hosting services. Multiple accusations of infringement could result in their users (and their whole households) being cut off from the Internet entirely.

Other countries have experimented with "three strikes" or "graduated response" systems like this, but ultimately abandoned them. Such measures unfairly deprive users of a basic necessity while failing to deter infringement.

### Any Good News?

U.S.-based OSPs may be relieved that the 512 Study did not recommend that Congress adopt the notice-and-staydown mandate that major copyright industry groups wanted and that the E.U.'s new strict liability rules now require for OSPs that host "large" amounts of user-uploaded content. Small and medium-sized OSPs, as well as nonprofit services, would find it difficult or impossible to take on the increased burdens and legal risks of notice-and-staydown regimes which require deployment of automated content recognition technologies.

Nor did the Study endorse empowering U.S. courts to issue no-fault site-blocking injunctions. Such injunctions would require Internet access providers to ensure their users cannot access offshore sites that host "pirat-

## Copyright law is supposed to promote the public good, not just maximize revenues for copyright industries.

ed" works, such as movies and sound recordings. No-fault injunctions raise serious due process and free speech concerns and could be abused because foreign sites that host non-infringing content cannot easily defend themselves in US courts.

### Striking Omissions

One striking omission is the Study's abject failure to recognize the interests of millions of individual user-creators who post their works on YouTube, Instagram, Twitter, Etsy, and myriad other platforms. They are authors too and their creations are entitled to copyright protection. These creators depend on hosting sites to share and/or commercialize their creations. The views of these creators about OSP liability rules should be given due recognition and weight. It is as if the Copyright Office considers conventional copyright industries as the only rights holders whose views about the safe harbors count.

Another striking omission is the Study's failure to consider that Internet-based companies are major contributors to the U.S. economy. An Internet Association-commissioned study reported this sector contributed more than \$2 trillion to the U.S. gross domestic product (GDP) in 2018, representing about 10% of GDP, and directly created six million jobs and indirectly supported an additional 13 million jobs. Most OSPs, moreover, experience few infringement claims, so major changes to the safe harbors will impose substantial costs on "good" actors without appreciably reducing infringements.

A third striking omission was the Study's failure to acknowledge that major copyright industries are thriving

in the digital age. One 2019 report on the state of the entertainment industry concluded that "the Internet, as currently structured, has been a creative force. It has helped many more people become creators and to make money from their creations, and the many industry sectors around 'copyright' are all seeing the fruits of that now."<sup>1</sup>

Perhaps most striking, though, was the Study's failure to give any weight to the interests of hundreds of millions of American Internet users who rely on this medium for a wide range of purposes: to get news, find information, share photos on social media, communicate with friends, family, and co-workers, seek entertainment, order consumer goods or food to be delivered, and in this pandemic age, to work safely from home and to educate our children when schools and libraries are closed. Copyright law is supposed to promote the public good, not just maximize revenues for copyright industries.

### Conclusion

The U.S. Congress is unlikely to make any changes to the DMCA safe harbors in the near term, especially given the coronavirus pandemic and the presidential election. The Copyright Office's 512 Study signals, however, the beginning of the next round of a prolonged battle in the halls of Congress over which, if any, of the many lopsided changes recommended in that report will ultimately be enacted into law. In the meantime, copyright industries will surely find the Study's analysis and conclusions useful to buttress their legal arguments in litigations against OSPs and to send more questionable takedown notices. **C**

### References

1. Masnick, M. and Beadon, L. *The Sky Is Rising: A Detailed Look at the State of the Entertainment Industry 41-42* (2019 ed.); <https://bit.ly/32zg0Rj>
2. U.S. Copyright Office, Section 512 of Title 17: A Report of the Register of Copyrights (May 2020), <https://bit.ly/3ki39c1>
3. Written Testimony of Jonathan Berroya, President and CEO, Internet Association, Hearing of the Senate Judiciary Committee, Subcommittee on Intellectual Property, Is the DMCA's Notice-and-Takedown System Working in the 21<sup>st</sup> Century? (June 2, 2020); <https://bit.ly/2Rpd6bp>

**Pamela Samuelson** (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley, CA, USA.

Copyright held by author.

## Economic and Business Dimensions Using Data and Respecting Users

*Three technical and legal approaches that create value from data and foster user trust.*

**T**RANSACTION DATA IS like a friendship tie: both parties must respect the relationship and if one party exploits it the relationship sours. As data becomes increasingly valuable, firms must take care not to exploit their users or they will sour their ties. Ethical uses of data cover a spectrum: at one end, using patient data in healthcare to cure patients is little cause for concern. At the other end, selling data to third parties who exploit users is serious cause for concern.<sup>2</sup> Between these two extremes lies a vast gray area where firms need better ways to frame data risks and rewards in order to make better legal and ethical choices. This column provides a simple framework and three ways to respectfully improve data use.

### Protecting Trust

Trust is a business asset. If you borrow against it, you can quickly become overdrawn. Earning consumer trust requires you to consider:

- ▶ How will you secure users' data?
- ▶ How will you ensure your product is reliable?
- ▶ How will you protect users' legal rights?
- ▶ How will you ensure data is collected and used ethically?

Users' legal rights include privacy, confidentiality, intellectual property, and contract details often found in the terms of service. Laws governing these



rights are fact-specific, vary by geography, and often in flux. Yet, even if the law permits you to use data in certain ways, should you? Ethical misuses, which may be legal uses, are often hidden from users and difficult to police. When three media outlets simultaneously reported Facebook's ethical missteps, the Cambridge Analytica scandal stripped more than \$100B from Facebook's value.<sup>a</sup>

### Risk Reduction Framework

One simple way to reduce data risk is to take the customer's perspective. Reducing risk means asking:

- ▶ Will data use meet the customer's expectations?

- ▶ Will they receive fair value in exchange for their data?

- ▶ Will they understand how their data is used?

- ▶ Will they have choice and control, even after their data has been sold?

- ▶ How would a firm's use of customer data play out in the court of public opinion; does this differ by country or culture?

Using the customer's perspective to place use-of-data cases on a heat map of reward-versus-risk suggests ethical considerations as shown in the figure on p. 29.

Evaluation starts from the *perspective of the customer who provides data*—not the business who collects it, nor other users, and certainly not third parties. Ethical and legal risks rise as perspective shifts away from the user.

<sup>a</sup> See <https://bit.ly/3bW0Fx9>

Risk also rises as data use shifts from a primary to a secondary purpose. A “primary” purpose is that for which the customer originally provided data. A “secondary” purpose is using the data for something else. Pregnancy apps are a great example. They collect extremely personal data and use it to deliver high-stakes insights. Providing a user with predictions on the days they might ovulate would be a primary purpose. Packaging their data with that of co-workers and selling it to employers or insurance companies to project maternity costs would be a secondary purpose.<sup>b</sup>

The more personal the data, the greater the risk to the firm and the consumer. In the green low-risk quadrant, anonymized data could deliver value to all users. A music-streaming app might analyze the size of customer files and speed at which they travel to improve performance for everyone. Risk rises if analysis touches content the customer considers confidential or when one firm fears data leakage to competitors. For example, competing services Spotify and Pandora might contract with the same cloud provider, who mines their content for analytic insights. A problem then arises if Pandora gets insights from Spotify data. To maintain trust and reduce risk, data analysis must give each data source full transparency about how a service works together with a compelling value proposition.

Given this framework, here are three ways to improve the reward-to-risk ratio.

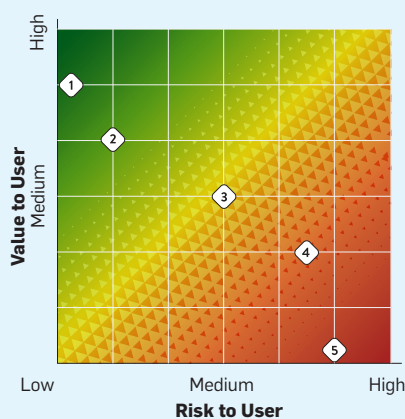
### Design for Value, Share with Users

Designing for user value expresses the obvious rubric: create more benefit than cost. Users are more or less willing to share data based on whether you give or take value. The same person might happily share a résumé that leads to a job opportunity but actively withhold that résumé if it were used for psychographic profiling and voter manipulation. Willingness to share data depends on *how* it is used and *who* gets the benefits. The ‘how’ should be ethical and the ‘who’ should emphasize the sharer. Design enters this calculation as it affects both parameters. One story from a grocer and one from an advertiser illustrate the shift in mind-set from third party to data source.

Groceries are a low-margin business, leading most grocers to sell customer loyalty data to third parties or use it for price discrimination. This creates little customer value and identifies the most price sensitive buyers. To address this challenge, one brand loyalty expert proposed a solution for a New England grocer. The new policy would use loyalty data to *protect* consumers. It would identify products with sugar, MSG, gluten, and peanuts and flag these on behalf of diabetics, celiacs, and people allergic to peanuts. This would decrease sales on flagged products and anger certain distributors. But, as a consumer, imagine your loyalty to a grocer who protects you from bloating, nausea, or diarrhea. Is it worth a price premium to be actively protected from harm? Under a protect-the-user policy, consumers may actively volunteer information to receive this value. Protecting customers increases both their willingness to participate and their willingness to pay. It shifts a grocer from low margins to loyal sales.

A second story concerns a ratings agency that tracks TV ad views to help networks price advertising. Concerned that viewers were skipping ads, the ratings agency designed ad-tracking and motion-sensing technology to learn what viewers saw at each instant. However, it was tone deaf to customer value. Even when paid, few viewers wanted spy systems in their homes just so third parties could learn about their private lives and sell ads.<sup>c</sup> A redesign focused on a mutually beneficial relationship. First, users gained control and could turn the system off. Second, repurposed motion sensors provided free home security and fire protection. These features compared favorably to less sophisticated systems that cost over \$30 per month. Although not yet fully deployed, a more sophisticated version could track “senior moments” and help trace likely locations of mislaid keys, glasses, and phones. Third, dashboards let users see their habits as well as any TV network could and manage the results. User-centered design provided transparency, choice, control, and fair value exchange. Ironically, J. Edgar Hoover used FBI spy systems to develop

### User Value vs. User Risk: North-West choices are safer than South-East choices.



1. Using the customer's data to fulfill their request
2. Using anonymized metadata for product improvement to benefit all users
3. Using the customer's data to personalize their service
4. Using the customer's data to make recommendations to other users
5. Selling or sharing the customer's data to 3<sup>rd</sup> parties for the benefit of 3<sup>rd</sup> parties

secret citizen files and harass political activists leading to public outcry in the 1950s and 1960s,<sup>4</sup> yet today Amazon and Google have sold more than 98 million home-listening devices in exchange for data on sports, news, weather, and users' personal calendars.<sup>d</sup>

### Save the Data, Discard the Detail

A second approach balances analytic flexibility with privacy. This method hinges on the insight that delivering value from data need not require access to *raw* data. Masked data, which cannot be converted back to its original form or linked to its source, can still permit analysis and even allow researchers to later ask unanticipated questions. Masked content goes beyond masked identity.

One such algorithm works by balancing two competing properties. The first step transmutes and reduces total available data; the second step aggregates sources. The first step represents lossy compression, where inessential entropy is discarded. Hashing represents one example. In the case of text, this step systematically makes individual words difficult to reconstruct by

<sup>b</sup> See <https://wapo.st/3moRlqb>

<sup>c</sup> A competitor that did this without consent got sued: <https://bit.ly/2ZBsoOF>.

<sup>d</sup> Cumulative sales since 2016. Source: <https://bit.ly/33u9ryl>



using morphological properties of language to shed linguistic detail while retaining root structure. It also discards enough information that subverting the algorithm via cryptanalysis becomes difficult.

The second step bundles masked information across individuals or across time in order to supply a corpus large enough to provide statistically meaningful pattern analysis. A more aggressive first stage provides greater privacy. A more aggressive second stage provides greater confidence in data analysis. To add protection, use lossier compression. To recover statistical power, aggregate more samples.<sup>e</sup> Individuals and individual messages become more difficult to read but populations and patterns get easier to resolve.<sup>3</sup>

Researchers used this method to analyze the relationships among email habits, content, and productivity of white-collar workers; yet no researcher could read any email involved in the study. Managers wanted to know, for example, ‘Does social network centrality predict productivity?’—yes. ‘Is communications diversity associated with productivity?’—yes, but with an inverted-U shape. More content diversity predicts revenues up to a point past which it implies lack of focus.<sup>1</sup> Using this technique, one could ask new questions to understand information diffusion, network diversity, responsiveness, content overlap, or even ad word targeting without reading literal content. Analysis of masked geolocation data or numbers could proceed analogously.

Of course, data masking must avoid infringing intellectual property rights and protect users’ other legal rights but keeping only masked data has three major benefits. It boosts willingness to share data. It reduces recording bias from users modifying their behavior. Most importantly, it reduces users’ risks even in cases of firms complying with the process of legal discovery or suffering a data breach.

### Save the Algorithm, Discard the Data

A third approach uses any number of

e There are key trade-offs. See Li, N., Li, T., and Venkatasubramanian, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of the IEEE 23rd International Conference on Data Engineering* (Apr. 2007), 106–115.

## An advantage of saving the model and discarding the data is that training on complete data can create models with great accuracy.

machine learning algorithms—neural networks, regression, random forests, *k*-means clustering, naïve Bayes, and so forth—to build a model of the world; then it saves that model but discards the data. Using this method, no data exists that could later be breached, compromised, de-anonymized, sold, or stolen yet it remains possible to classify a new image or to predict a new product’s popularity. Another method, secure multiparty computation (MPC), splits the data among several independent parties. Each party can perform calculations on their partition but not see how the results combine. A third party combines results but cannot see the data. This limits access to data during the same calculation whereas discarding data limits access in future calculations.<sup>f</sup>

An advantage of saving the model and discarding the data is that training on complete data can create models with great accuracy. The AlphaGo machine learning algorithm beat the world’s expert at the game of GO.<sup>g</sup> A different algorithm beat human lawyers at analyzing risks present in non-disclosure agreements (NDAs).<sup>h</sup> A third algorithm predicts the onset of strokes and heart attacks more accurately than doctors.<sup>i</sup> Another detects breast cancers with 99% accuracy.<sup>j</sup> The disadvantage of finely tuned machine learning models is that they cannot be used for purposes outside their training. You cannot get good answers to questions you did

f See <https://bit.ly/3kCm9ID>

g See <https://bit.ly/2E0Ubr3>

h See <https://bit.ly/35zBv67>

i See <https://bit.ly/2ZDpB7B>

j See <https://bit.ly/2RnTu7k>

not ask. If raw data is gone, there is no retraining option. By contrast, the advantage of saving masked data as in the second point here—save the data, discard the detail—is that one can ask new questions that one overlooked initially. However, the disadvantage is that the loss of information causes model accuracy to fall relative to analysis of raw data.

Keeping only the final trained algorithm naturally limits future applications to a primary purpose—the one used to train the model. Using the model for a different purpose would require access to raw data for retraining. The absence of this data limits secondary uses, which limits legal and ethical risk.

### Conclusion

These three approaches—designing for user benefit, saving masked data, and saving masked algorithms—each improve a user’s reward-to-risk ratio. Design for user benefit increases the value to users and pushes points North on the figure heat map. Saving masked data and masked algorithms reduces user profiling, secondary uses, and third-party access, pushing points West in the figure. Together, these three approaches offer a range of ways to deliver value from data analysis while protecting users and respecting their trust. Approaching data analysis from the perspective of the user who provided data is not only good business and legal advice but also a way to strengthen ethics and relations with users. ■

### References

1. Aral, S. and Van Alstyne, M. The diversity-bandwidth trade-off. *American Journal of Sociology* 117, 1 (Jan. 2011), 90–171.
2. Cadwalladr, C. ‘I made Steve Bannon’s psychological warfare tool’ Meet the data war whistleblower. *The Guardian* (Mar. 18, 2018).
3. Reynolds, M., Van Alstyne, M. and Aral, S. Privacy Preservation of Measurement Functions on Hashed Text Annual Security Conference. *Discourses in Security Assurance and Privacy*. (Las Vegas, NV, Apr. 15–16, 2009). Information Institute Publishing, 41–45.
4. Theoharis, A.G. and Cox, J.S. *The Boss: J. Edgar Hoover and the Great American Inquisition*. Temple University Press, Philadelphia, PA, 1988.

**Marshall W. Van Alstyne** ([mva@bu.edu](mailto:mva@bu.edu)) is a Questrom Chair Professor at Boston University where he teaches information economics. He is also a Digital Fellow at the MIT Initiative on the Digital Economy and co-author of the international best-seller *Platform Revolution*. W.W. Norton, 2016.

**Alisa Lenart** ([Alisa.lenart@autodesk.com](mailto:Alisa.lenart@autodesk.com)) is Senior Corporate Counsel, Cloud Platform, Autodesk, San Francisco, CA, USA.

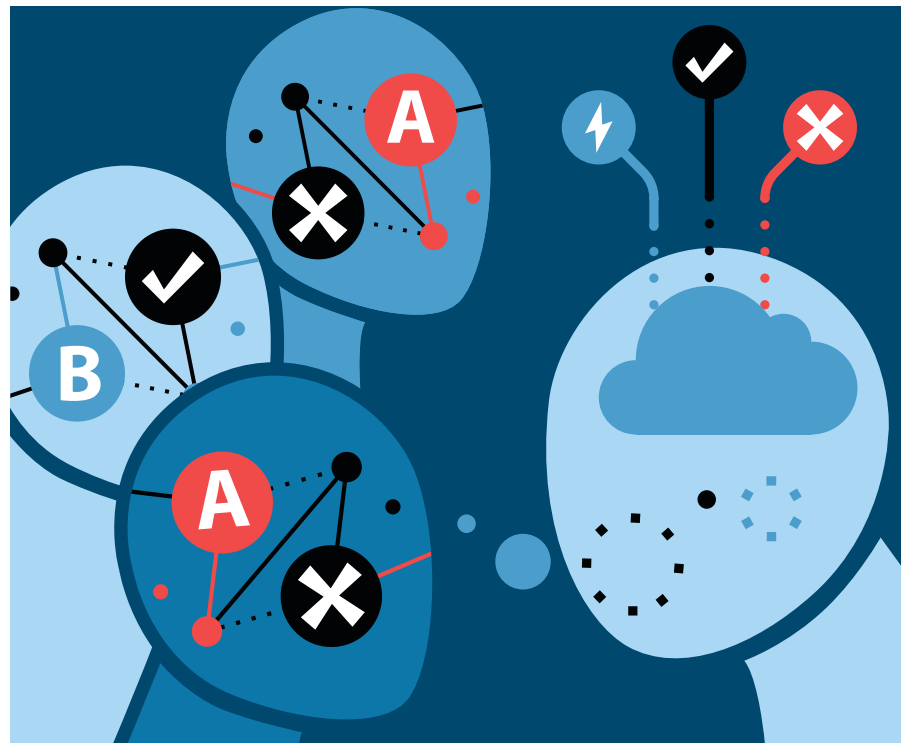
## Education

# It Is Time for More Critical CS Education

*By which 'critical' means an intellectual stance of skepticism, centering the consequences, limitations, and unjust impacts of computing in society.*

**W**E LIVE IN uncertain times. A global pandemic has disrupted our lives. Our broken economies are rapidly restructuring. Climate change looms, disinformation abounds, and war, as ever, hangs over the lives of millions. And at the heart of every global crisis are the chronically underserved, marginalized, oppressed, and persecuted, who are often the first to befall the tragedies of social, economic, environmental, and technological change.<sup>3</sup>

You might think these issues have little to do with computing. But you would be wrong. The weaving of computing through society has not only involved computing in these crises, but, in many ways, placed computing at their centers. Computers increasingly mediate our communication. Automation is accelerating economic restructuring, destabilizing work, and devaluing labor. The demand for information is increasing carbon outputs and exploitative mining of rare metals. Social media is amplifying falsehoods. The Internet is the new battleground of modern warfare. And in all of these systems, data and algorithms amplify racism, sexism, heterosexism, ableism, ageism, xenophobia, cisheteronormativity, and other forms of inequity, injustice, and bias.<sup>2,3</sup> Computing does not occur in a vacuum: it shapes and is shaped by ever-evolving social, cultural, institutional, and political forces.



These links between computing and injustice seem invisible to many, including those who bear the brunt of these injustices. Some young people grow up seeing computers as magical machines that bring joy, escape, and connection. Others experience them as vectors for violence, sexual harassment, cyberbullying, addiction, and isolation. Some adults view computing as a force of economic growth and progress. Others experience subjugation to unjust algorithmic decisions about their loan eligibility, work

schedules, and unemployment insurance, yet lack the computing literacy to counter authoritative voices on these algorithms' designs. Meanwhile, many of us in the computing discipline, while happy to celebrate computing as a tool for social change,<sup>1</sup> ignore its role in these injustices,<sup>2</sup> and in some cases, dismiss the idea that computing is anything but a value-neutral tool independent from society.

We argue, as others have,<sup>5</sup> that making these injustices visible to society is the responsibility of CS educators.



Association for  
Computing Machinery

2018 JOURNAL IMPACT  
FACTOR: 6.131

## ACM Computing Surveys (CSUR)

*ACM Computing Surveys* (CSUR) publishes comprehensive, readable tutorials and survey papers that give guided tours through the literature and explain topics to those who seek to learn the basics of areas outside their specialties. These carefully planned and presented introductions are also an excellent way for professionals to develop perspectives on, and identify trends in, complex technologies.



For further information  
and to submit your  
manuscript,  
visit [csur.acm.org](http://csur.acm.org)

After all, educators hold the power to shape public perception of computing. We do this through the problems we focus on in our classrooms; through who we choose to teach; in how we shape students' career choices; and in how we conceptualize computing to journalists, social scientists, and society. The world has critical questions about computing and it is time we started teaching more critical answers.

While there are many ideas to teach, we believe three ideas are key.

### Computing Has Limits

Computing is powerful and the allure of this power is compelling. It is what drives students to our classrooms, it is what has led to worldwide calls for CS for All in primary and secondary schools, and it is what has made some of our lives better than ever, providing more information, connection, opportunity, and voice.

But the belief in computing's limitless power has led many of us to believe that computing *always* makes things better.<sup>1</sup> This could not be further from the truth. Judges, for example, have begun to delegate sentencing decisions to recidivism prediction software, ignorant of the racially biased data upon which those predictions are based. Our global climate agreements rest heavily upon the assumption that technology, and not behavior change, will save us from calamity. Investors have amplified the computing-enabled gig economy not because it is an inherently more humane form of human labor, but because it profits a small group of private investors and saves those with means and money a bit of time.

All of these troubling trends emerge from a set of neophilic myths: that software is always right, that software is always value-neutral, and that software can solve every problem. CS education must replace these conceptions with the reality that software is often wrong; software always embeds its creators' values and biases; and software can only solve some problems, and many cases, creates new ones.

### Data Has Limits

Computing has little value without data. People come to Facebook not

for the newsfeed algorithm, but for the content their friends and family write. People come to Google, Baidu, and Yandex not for ranking algorithms, but for the Web pages millions have carefully authored. People watch Netflix, iQIYI, and Tencent not for their recommendations, but for television, movies, and events. And while these algorithms are useful, their value is dependent on the quality of the data they process: imperfect, biased inputs lead to imperfect, biased outputs.<sup>3</sup>

But computing often subordinates data, ignoring the cost of creating it, the individuals and social contexts from which it is wrought, and its role in global crises and injustices. After all, it is the desire for data that drives the carbon output of datacenters; it is biased datasets that enable facial recognition algorithms to work so well for white people, subjecting everyone else to greater risk of accidental prosecution by automated surveillance; and it is binary classifications in airport security scanners that, trained on cisnormative bodies, cause trans and non-binary people to be physically harassed for "bodily anomalies."<sup>2</sup> Data is responsible for many harms of computing, whether directly through its collection or indirectly through its use.

Thus, all CS educators must teach what information science and librarians have long known: data is always about the past and not the future; data is always an imperfect and biased record, encoding the values, beliefs, and ideas of its creators; and incorrect interpretations and uses of data harm people in unequal ways.<sup>4</sup>

**Data is responsible  
for many harms  
of computing,  
whether directly  
through its collection  
or indirectly  
through its use.**



## CS Has Responsibility

Many early conceptions of CS education view computing as a medium for expression. And this view has dominated: we celebrate what students and companies create, partly in recognition of the inherent difficulty of programming. But while we give great attention to how our students create things, and the scale of impact their creations have on the world, we often leave the moral choice about what to create to individuals and investors.

However, the choices that developers make when they create with computing are not purely individual or capitalistic. They are inherently social and collective, and infused with value judgments. For example, when a CS graduate accepts their first job, they are endorsing and investing in the values of the company they choose; students should be supported in reflecting on this endorsement. Similarly, when engineers at Google internally protested the creation of a censored search engine for China, they were doing it on behalf of not only themselves, but China and the rest of the world.

CS education at all levels must center these responsibilities and value tensions, ensuring all people—not just CS majors—understand that creating software comes with collective responsibilities to society.

## Ways Forward

Many respond to these concerns by advocating for everyone to learn to code, arguing that programming forces us to confront the limitations of computing, the necessity of data, and the role of programmers in shaping software. But learning to code often leads people to view programs as powerful rather than perilous, data as abstract and free of bias, and programmers as clever wizards rather than social actors. And yet, more people know how to code than ever, and critical views on computing are still rare in CS education and industry.

What will make them more common? An intentional effort to develop a critical literacy of computing, helping everyone understand the social and cultural systems that drive computing, and the social and cultural systems disrupted by computing. This

## Realizing a more critical CS education requires more than just teachers: it also requires CS education research.

means educating primary, secondary, and post-secondary CS teachers who can help everyone see computing as both a powerful medium for expression *and* a perilous tool for oppression. It means preparing CS teachers who can develop students' sense of collective civic responsibility. And it means more than just an ethics requirement for CS majors: it means recasting computing itself in moral, ethical, and social terms.

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? And how do we prepare outstanding primary, secondary, and post-secondary teachers to equitably teach these ideas to everyone in a way that is responsive to local needs and values?

If we can answer these research questions and enact their implications in our teaching, we may see students create (and demand) a more inclusive future for computing. We may see social media stabilize free press and democracy rather than supplant it. We may see a generation of students choose to invest their skills in broader global problems of healthcare, energy, education, and government. And we might see a more just use of algorithms and machine learning.

Work on these futures has only just begun. Researchers around the world are shifting their attention to algorithmic fairness, data bias, and CS ethics education. Grassroots communities are advancing design justice, critically analyzing the role of computing in society.<sup>2</sup> Even ACM's own Future of Computing Academy, which periodically brings together new computing faculty to envision the discipline, recently called not for more innovation, influence, or impact in CS, but more humility. These grassroots movements outside of computing, and our own nascent conversations within computing, inspire some hope. Now it is time to translate that hope into more critical CS education. ■

### References

1. Ames, M.G. *The Charisma Machine: The Life, Death, and Legacy of One Laptop per Child*. MIT Press, 2019.
2. Costanza-Chock, S. *Design Justice: Community-led Practices to Build the Worlds We Need*. MIT Press, 2020.
3. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.
4. Rubin, A. Learning to reason with data: How did we get here and what do we know? *Journal of the Learning Sciences* 29, 1 (2020).
5. Vakil, S. Ethics, identity, and political vision: Toward a justice-centered approach to equity in computer science education. *Harvard Educational Review* 88, 1 (2018).

**Amy J. Ko** (ajko@uw.edu) is a professor in The Information School, University of Washington, Seattle, WA, USA.

**Alannah Oleson** (olesona@uw.edu) is a Ph.D. student in The Information School, University of Washington, Seattle, WA, USA.

**Neil Ryan** (neilryan@cs.washington.edu) is a Ph.D. student in The Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA.

**Yim Register** (yreg@uw.edu) is a Ph.D. student in The Information School, University of Washington, Seattle, WA, USA.

**Benjamin Xie** (bxie@uw.edu) is a Ph.D. student in The Information School, University of Washington, Seattle, WA, USA.

**Mina Tari** (minatari@uw.edu) is a Ph.D. student in The Information School, University of Washington, Seattle, WA, USA.

**Matthew Davidson** (mattjd@uw.edu) is a Ph.D. student in The College of Education, University of Washington, Seattle, WA, USA.

**Stefania Druga** (st3f@uw.edu) is a Ph.D. student in The Information School, University of Washington, Seattle, WA, USA.

**Dastyni Loksa** (dloksa@towson.edu) is an assistant professor at Towson University, Towson, MD, USA.

This column is derived from the first author's 2019 keynote at the ACM Koli Calling conference, available at <http://faculty.uw.edu/ajko/talks>

Copyright held by authors.

## Viewpoint

# Where Should Your IT Constraint Be? The Case of the Financial Services Industry

*Locating the strategic location of the IT function constraint.*

**T**HE FINANCIAL SERVICES industry is in a state of turmoil due to technological innovation and the industry's move from brick-and-mortar to digital and particularly mobile service delivery. In most cases, information technology (IT) function is the constraint blocking the organization from innovating rapidly.

In this Viewpoint, we seek to locate where the constraint of the IT function should strategically be. In recent work<sup>2-6</sup> the focus was on the tactical management of a system using the theory-of-constraints (TOC). Research done so far has not addressed the strategic location of the constraint.

Here, we focus on the financial services industry. Though digitalization is applicable to a multitude of industries, we focus on the financial services industry due to its particular characteristics: dramatic reduction of human resources, significant increase of self-service, end-to-end computerization due to the lack of physical entities, increasing IT budgets, and a similarity of problems between banks, insurance, and credit card companies.

### Recent Tactical Work

The theory-of-constraints<sup>3</sup> has developed a methodology to identify the sys-



tem's current constraint and manage it, commonly using the seven focusing steps (see Figure 1) as the tool for this purpose.<sup>6</sup>

The seven steps set the organization's goal and performance measures, identify the current system constraint, exploit the constraint, subordinate other functions to the constraint and ele-

vate (offload) the constraint. If a constraint is broken, attention is turned to the new constraint. These techniques have been documented to increase the organization's throughput, significantly improving its performance. This Viewpoint deals with the strategic management of the constraint, addressing the question of where the constraint

should be. A common graphical tool to identify the system's constraint is the Cost-Utilization analysis,<sup>6</sup> which looks at the cost of each organizational function and at the extent to which it is utilized. Functions that are fully utilized are the organization's bottlenecks. The constraint should be located in the most expensive/scarce resource.<sup>6</sup>

Figure 2 presents the Cost-Utilization diagram of the financial industry. Previous research<sup>5</sup> has shown that IT is a permanent bottleneck, because demand for IT services is X3 to X5 the capacity of the IT function and because of the scarcity of good IT professionals and partners. Here, "IT function" applies to supporting the whole software life cycle. As seen in Figure 2, the cost (represented by the width of the horizontal axis) is large and the utilization is 100%. Later on, we will drill down to identify specifically where in the IT function the constraint should be located. Demand is the result of the needs of all organizational functions (marketing, operations, sales, financial markets, and so forth), with compliance, maintenance of existing systems, change requests, fix-it jobs, technological project requirement, and so forth, accounting for a significant portion of the demand.

As shown in Figure 2, sales is a permanent bottleneck because demand is endless (upsell, cross-sell, new customers and so forth) Figure 2 illustrates that other functions such as operations, logistics, legal services, and others, should never be the organization's constraint.

A current-reality-tree (CRT) of the financial services industry is presented in Grosfeld-Nir et al.;<sup>4</sup> this is based on Goldratt's work.<sup>2</sup> For this Viewpoint, we analyzed the phenomenon that "Insufficient value is generated through digitalization" and drew the fCRT (focused-Current-Reality-Tree), a lighter version of Goldratt's CRT6 (see Figure 3).

The fCRT presents the leading Undesirable-Effect (UDE) that insufficient value generated through digitalization. All other UDEs account directly or via other cause-and-effect relationships to the leading UDE and stem from the root-causes: "Not enough differentiation of users" and "No clear identification of constraint location". Here, we deal with these root causes.

Figure 1. The seven focusing steps.

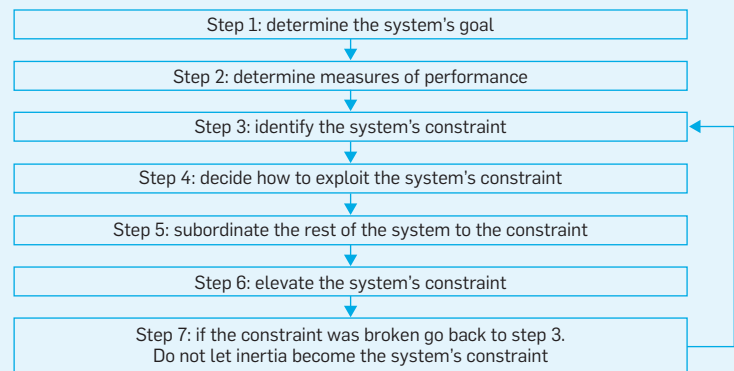


Figure 2. Cost-Utilization diagram of the financial industry.

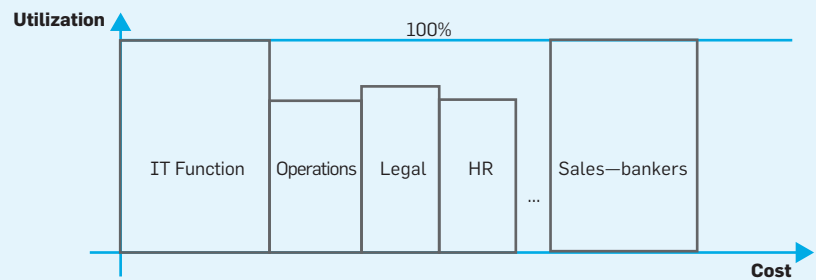
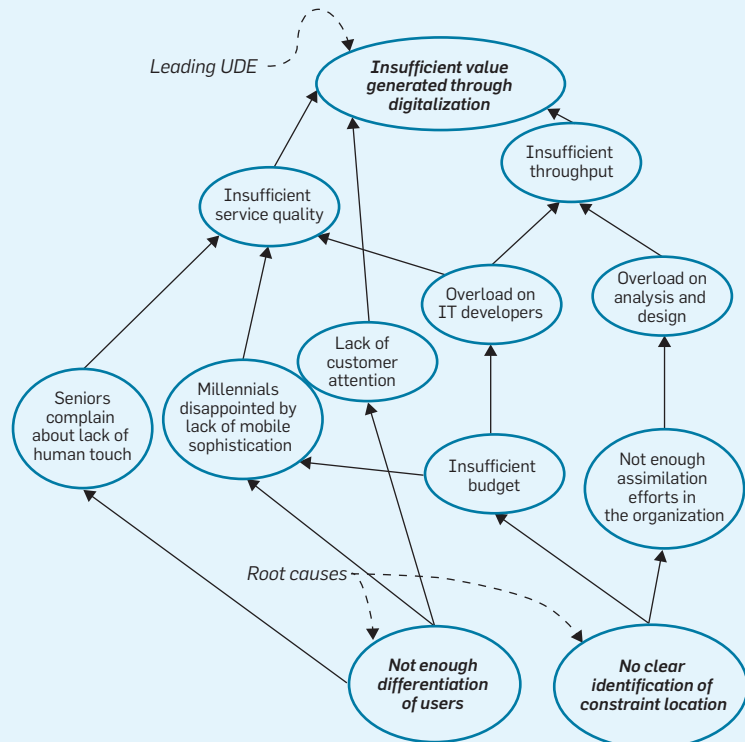
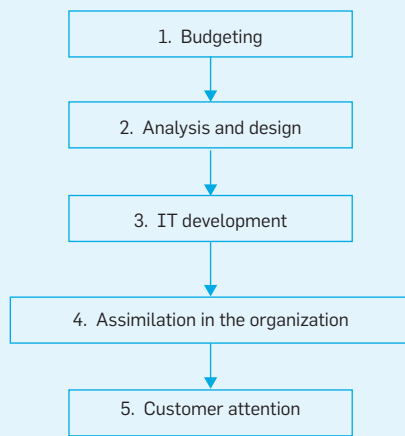


Figure 3. Focused Current Reality Tree.





**Figure 4. Constraint positioning ‘immediate suspects’.**

### Constraint Positioning “Immediate Suspects”

Given the digitalization flow, we now search for potential resources constraining IT digitalization value creation. Five ‘immediate suspects’ are identified: Budgeting, Analysis and design, IT development, Assimilation in the organization, and Customer attention. Figure 4 describes bottleneck positioning ‘immediate suspects’. These are plausible constraints:

**1. Budgeting.** Managers will always complain the budget is insufficient. This is often the case, but we have encountered a dozen large financial institutions that did not utilize all of their budgets by year-end. The financial mind-set is that if you provide enough of it, money will solve resource shortage. In reality, the pace of growth

is limited. The budgeting process should be divided into two steps: first, how much funding is management and the board of directors willing/capable to allocate, given large, multi-year infrastructure projects; and second, how much to allocate to each department/project.

**2. Analysis and design.** In many organizations, the system analysts are part of the business-relationship-management (BRM) function, and constrain new project development. It is not cost effective that analysts and designers should be the system’s constraint. Imagine the board of directors’ reaction should the CIO argue that projects are late due to the shortage of 10 analysts or designers. Despite their scarcity, their numbers are not large and their cost is manageable. This problem is solvable within the budget.

**3. IT Development.** Development resources are expensive (whether internal or outsourced) and often block development throughput. Once the budget is allocated, projects should go through a value-based strategic gating process, eliminating projects that do not generate enough value or return-on-investment (ROI). In some cases, with a positive ROI management is unwilling to increase its IT headcount or budget. From our experience, the throughput of the IT development function can be significantly improved by using tactical methodologies such as the complete-kit concept, TOC, tactical gating, proper measurement, introduction

of DevOps tools, and so forth.<sup>5</sup>

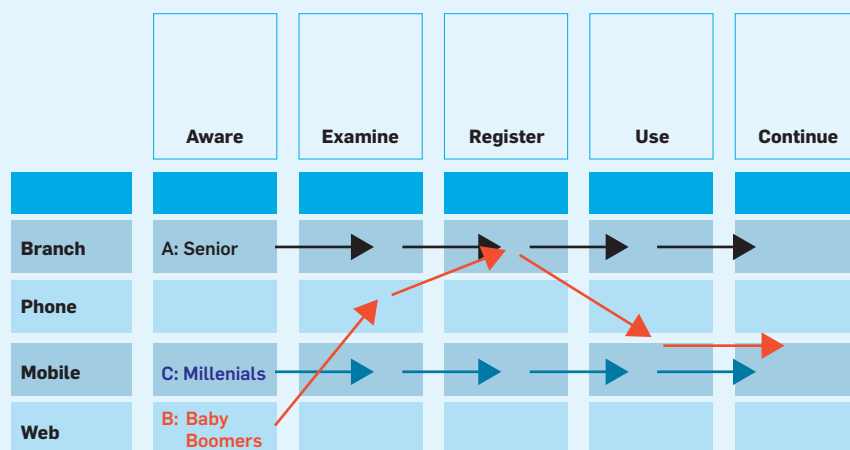
**4. Assimilation in the organization.** There is a limit to the amount of new systems the organization can absorb at a given time (see Davenport and Beck<sup>1</sup>). We have not encountered an instance where assimilation was the constraint and the whole organization was subordinated to it. Assimilation should be planned and staggered over time. Agile methodology can be used to sequentially launch features in an effective pace. The introduction of too many systems at a time results in chaos and bad multitasking. Thus, application of scheduling tools can smooth the load on the organization.

**5. Customer attention.** In today’s disruptive arena, many customers fail to become familiar with, and use, available applications introduced by the organization. More systems/functions are introduced than the customer is capable of adopting. The customer’s journey illustrates customer personas aware of a new digital service, examining it, registering for it, actually using it and continuing doing so. Figure 5 illustrates the journeys of three personas: seniors (persona A), performing their interactions physically in the branch; baby boomers (persona B), searching the Web, examining by phone, being coached in the branch and then using the mobile app.; millennials (persona C) performing all journey phases via their mobile phones (see Figure 5).

Pitfalls and hurdles along the way result in customer usage churn. Consider the paradox that after spending scarce, expensive capital, effort and budgeting the IT department, solving specification bottlenecks, screening projects in the development department, improving the development department, solving assimilation problems, and then—being blocked on the customer attention bottleneck; we next analyze the three personas in terms of their attention and usage profile.

### Analysis of Persona Resource Usage Profile

Pareto analysis is one of the most powerful tools in management stating that 20% of the phenomena account for 80% of the effect, also known as the 80/20 rule. The Pareto methodology is

**Figure 5. Customer journeys.**

an extension of the Pareto rule stating that improving a system is accomplished by the following three steps: classification, differentiation, and resource allocation.<sup>4</sup> Dealing with customer attention, we can apply this methodology as follows:

**Step 1: Classification.** We classify the customers as follows: “A” customers are 30% of the customers that account for 70% of face-to-face and phone support. These are mostly senior customers unaccustomed with many technologies. “B” customers are 40% of the customers that account for 20% of face-to-face and phone support. These are mostly baby boomers who will use digital services if provided with friendly assistance. “C” customers are 30% of the population that account for 10% of face-to-face and telephone traffic. They are technically savvy millennials (who would rather go to the dentist than visit a bank<sup>a</sup>). They can handle it alone, needing occasional assistance via Google, chat, or videos.

**Step 2: Differentiation.** Adopting a different policy for each group. No amount of effort will convert “A” customers to digital. Our challenge is to shift them to the telephone channel. Branch personnel should patiently train them to use the telephone to reduce the load on the branch. They are not a target for digital banking, and no resources are allocated to convert them. They are therefore not a constraint. The attention span of “B” customers limits their adoption of new technologies. They are the focus of technology introduction as they account for the lion’s share of our target population (estimated 40% of the population that account for merely 20% of the face-to-face traffic). “C” customers crave digital services, can assimilate all the changes and are not a constraint.

**Step 3: Resource Allocation.** “A” customers are allocated telephone services delivering basic bread-and-butter services and dedicated staff in the branch training them to use these services. “B” customers are allocated technology facilitation resources and are the focal point of the digital world. “C” customers are provided

<sup>a</sup> See <https://bit.ly/2Rgf4L8>

## The IT-development potential bottleneck can be resolved by a value-based strategic gating process and applying tactical tools such as the complete-kit concept and others.

with virtually paperless mobile banking services.

### Discussion

In our journey, we have analyzed several ‘immediate suspects’: budgeting, analysis and design, IT development, assimilation in the organization, and customer attention. Analysis and design is not the constraint as reasonable investment can resolve the problem. The IT-development potential bottleneck can be resolved by a value-based strategic gating process and applying tactical tools such as the complete-kit concept and others, and sometimes budget increment. Assimilation in the organization should not be the bottleneck and can be resolved by smoothing the peak demand. Customer attention should not be the bottleneck as only B population requires organizational resources to adopt new technologies.

Traveling through potential bottlenecks has brought us to the realization the budget should be the constraint. However, management should ensure:

- ▶ the budget is allocated through a structured and value-based strategic gating process;
- ▶ analysis and design is budgeted for and should not be a bottleneck;
- ▶ development must undergo drastic improvement that will reveal capacity to the system;
- ▶ assimilation should not be the bottleneck, and if it is, then management should implement smoothing of feature launch; and

- ▶ from all personas only the “B” population (baby boomers) consume system resources, they are not considered the constraint since they constitute only 20% of the population.

### Conclusion

There is a set of issues regarding the strategic positioning of a constraint. Managers should resolve the three following issues:

- ▶ Where the constraint should be.
- ▶ Where the constraint is now.
- ▶ How to move the constraint to the desired position.

Analyzing the IT development process, backed by well-accepted methodologies such as TOC and Pareto analysis, we have come to the conclusion that the budget should be the constraint. Executives must generate a well thought-out budget and at the same time ensure other functions do not become the system’s constraint, according to the guidelines here.

IT executives have a significant message to convey: the budget is the constraint. All other issues must be resolved in light of this constraint. The board of directors and management define the budget and the role of the IT executives is to participate in the budgeting process, resolve problems, and lead the analysts and developers according to the budget.

Similar analysis of other industries such as telecom, retail, tourism, and so forth and may yield other results. A company that follows these guidelines will better compete in the new market. ■

### References

1. Davenport, T.H. and Beck, J.C. *The Attention Economy: Understanding the New Currency of Business*. Harvard Business School Press, Boston, MA, 2001.
2. Goldratt, E.M. *It’s Not Luck*. North River Press, Croton on Hudson, NY, 1994.
3. Goldratt, E.M. and Cox, J. *The Goal: A Process of Ongoing Improvement* (2<sup>nd</sup> ed.). North River Press, Croton on Hudson, NY, 1986.
4. Grosfeld-Nir, A., Ronen, B. and Kozlovsky, N. The Pareto managerial principle: When does it apply? *International Journal of Production Research* 45, 10 (Oct. 2007), 2317–2325.
5. Pass, S. and Ronen, B. Reducing the software value gap. *Commun. ACM* 57, 5 (May 2014), 80–87.
6. Ronen, B. and Pass S. *Focused Operations Management*. John Wiley and Sons, Inc. Hoboken, NJ, 2008.

**Boaz Ronen** (boazr@tauex.tau.ac.il) is a Professor Emeritus at The Coller School of Management, Tel Aviv University, Israel.

**Alex Coman** (coman@mta.ac.il) is a Senior Lecturer at The Academic College of Tel Aviv-Yaffo, Israel.

## Viewpoint

# Reason-Checking Fake News

*Using argument technology to strengthen critical literacy skills for assessing media reports.*

**W**HILE DELIBERATE MISINFORMATION and deception are by no means new societal phenomena, the recent rise of fake news<sup>5</sup> and information silos<sup>2</sup> has become a growing international concern, with politicians, governments and media organizations regularly lamenting the issue. A remedy to this situation, we argue, could be found in using technology to empower people's ability to critically assess the quality of information, reasoning, and argumentation through technological means. Recent empirical findings suggest "false news spreads more than the truth because humans, not robots, are more likely to spread it."<sup>10</sup> Thus, instead of continuing to focus on ways of limiting the efficacy of bots, educating human users to better recognize fake news stories could prove more effective in mitigating the potentially devastating social impact misinformation poses. While technology certainly contributes to the distribution of fake news and similar attacks on reasonable decision-making and debate, we posit that technology—argument technology in particular—can equally be employed to counterbalance these deliberately misleading or outright false reports made to look like genuine news.

### From Fact-Checking to Reason-Checking

The ability to properly assess the quality of premises and reasoning in persuasive or explanatory texts—critical



literacy—is a powerful tool in combating the problem posed by fake news. According to a 2017 Knight-Gallup survey, one in five U.S. adults feels “not too confident” or “not confident at all” in distinguishing fact from opinion in news reporting.<sup>a</sup> Similarly, in the U.K., the National Literacy Trust recently reported that one in five British children cannot properly distinguish between reliable online news sources and fake news, concluding that strengthening

a See <https://kng.ht/3iy9z6p>

critical literacy skills would help in identifying fake news.<sup>b</sup>

Efforts to combat the effects of fake news focus too often exclusively on the factual correctness of the information provided. To counter factually incorrect—or incomplete, or biased—news, a whole industry of fact-checkers has developed. While the truth of information that forms the basis of a news article is clearly of crucial importance, there is another, often overlooked, aspect to

b See <https://bit.ly/3mxjR9s>



fake news. Successfully recognizing fake news depends not only on understanding whether factual statements are true, but also on interpreting and critically assessing the reasoning and arguments provided in support of conclusions. It is, after all, very possible to produce fake news by starting from true factual statements and drawing false conclusions by applying skewed, biased, or otherwise defective reasoning. We therefore argue that fact-checking should be supplemented with reason-checking: evaluating whether the complete argumentative reasoning is acceptable, relevant, and sufficient.<sup>3</sup>

### Argument Technology for Critical Literacy

Seven years ago, we introduced the Argument Web in *Communications*: an integrated platform of resources and software for visualizing, analyzing, evaluating, or otherwise engaging with reasoned argument and debate.<sup>1</sup> Since then, argument technology has matured into an established research field, attracting widespread academic and industrial interest. Recently, for example, IBM presented the results of its ‘grand challenge’ on argument technology in a live debate between the Project Debater system and two human debating champions.<sup>2</sup>

Commissioned by the BBC, we have developed a suite of argument technologies, built on the infrastructure of the Argument Web. The resulting software tools are aimed at providing insight into argumentative debate, and at instilling the critical literacy skills needed to appraise reasoned persuasive and explanatory communication. In addition to identifying reasoning patterns and fallacies, our software addresses the issue posed by echo chambers in which people are less exposed to opinions diverging from their own, while already held views get reinforced. Several cognitive processes are involved in this process—such as confirmation bias<sup>6</sup>—which discourages the consideration of alternative positions in a dispute, and the backfire effect,<sup>8</sup> which leads to further entrenchment of viewpoints when presented with conflicting facts.

The Polemicist application<sup>d</sup> addresses this looming one-sidedness of argumentative positions. The application lets the user take on the role of moderator in a virtual radio debate: selecting topics, controlling the flow of the dialogue, and thus exploring issues from various angles. The textual data is drawn from the Argument Web database of analyzed episodes of BBC Radio 4’s Moral Maze.<sup>e</sup> On this weekly radio program, recurring panelists and invited subject experts debate a morally divisive current affairs topic. The ensuing debate is often lively, combative, and provocative, producing a wealth of intricate argumentative content. Polemicist produces responses given by the actual Moral Maze participants from the Argument Web database and assigns them to software agents modeled on the participants. Playing the role of moderator lets the user rearrange the arguments and create wholly novel virtual discussions between the contributions of participants that did not directly engage in the original debate, while still reflecting their stated opinions.

Test Your Argument<sup>f</sup> aims to both foster critical literacy skills and prompt users to consider alternative viewpoints. The software challenges users with a number of argumentation puzzles designed to help develop an understanding of the core principles of

strengthening and critiquing arguments. The examples are again drawn from debates on BBC Radio 4’s Moral Maze. Test Your Argument was launched on BBC Taster in December 2017, and has since been visited over 10,000 times, with a rating of 4/5, and 88% of the evaluations saying that the BBC should do more along these lines.<sup>g</sup>

Argument Analytics serves as an online second-screen supplement to BBC Radio and Television broadcasts. Tried in 2017 on selected episodes of Moral Maze, the data-driven infographics are designed to provide a deeper insight into the debate. For instance, the interaction between arguments pro and contra are diagrammatically visualized, alignment between participants’ stances is mapped out, and a timeline shows which parts of the debate lead to the most conflict. An example of Argument Analytics for the October 11, 2017 episode of Moral Maze dedicated to the 50-year anniversary of the Abortion Act in the U.K.<sup>h</sup>

### Argument Mining for Reason-Checking

The latest addition to the suite of argument technologies developed for the BBC is The Evidence Toolkit.<sup>i</sup> This online application is designed to encourage users to dissect and critically

d Online at <http://polemici.st>  
e See <https://bbc.in/3kqh7J2>  
f See <https://bbc.in/3iDTRGw>

g All user statistics reported in this Viewpoint are obtained directly from the host and are correct at the time of writing.  
h See <http://bbc.arg.tech>.  
i Available at <https://bbc.in/2FFNQen>

#### The Evidence Toolkit interface.

The screenshot displays the Evidence Toolkit interface. At the top, it shows the title 'The Evidence Toolkit' and the source 'BBC'. The main content is a news article titled 'Air pollution: Are diesel cars always the biggest health hazard?' with a sub-headline 'Diesel engines produce higher levels of particulate emissions than petrol engines, but they also help reduce carbon dioxide emissions'. The article text is partially obscured by several analysis tools and filters. On the left, there's a 'HELP' section with a 'CRITICAL THINKING' icon. On the right, there's a 'PARTICULATE FILTERS' section with a 'PARTICULATE FILTERS' icon. At the bottom right, there's a 'REASON CHECKER' section with a 'REASON CHECKER' icon. The interface also features a 'BBC' logo and a 'This Page' link.

c See <https://ibm.co/2Rzb7RW>

appraise the internal reasoning structure of news reports. The Evidence Toolkit launched in March 2018<sup>j</sup> as part of BBC's Young Reporter initiative (formerly called 'BBC School Report'). BBC Young Reporter is a U.K.-wide opportunity offered to some 60,000 11- to 18-year-old students to develop their media literacy skills by engaging firsthand with journalism and newsmaking.<sup>k</sup> The 2018 initiative addressed the issue of fake news. To let students develop the means to distinguish real news from fake news, the BBC commissioned the iReporter game<sup>l</sup> from Aardman Animations targeted at 11- to 15-year-olds, and The Evidence Toolkit from the Centre for Argument Technology<sup>m</sup> for 16- to 18-year-olds.

The Evidence Toolkit guides students through a series of steps to identify claims, arguments, counter-arguments, reasoning types, and evaluation criteria on the basis of scholarship in Critical Thinking and Argumentation Theory.<sup>3,9</sup> To help students understand the theoretical concepts, examples are given from episodes of BBC Radio 4's Moral Maze. Since BBC Young Reporter is intended to be primarily used in the classroom and teachers will often not be argumentation experts themselves, The Evidence Toolkit comes with teacher notes, available through the BBC website.<sup>n</sup>

Upon identifying the main claim and reasons in the news article, the software helps users classify the reasoning in the news article based on the type of evidence provided. Reasons can be connected to claims in many different ways. Drawing on theories of argumentation and persuasion,<sup>9</sup> users are presented with a compact set of options not requiring any specific theoretical background knowledge. Reasoning is classified as fact-based or opinion-based, which in turn can be subdivided further. Opinion-based reasoning, for example, subdivides into evidence drawn from experts (providing authoritative backing), from popular sentiment (of the masses or of a particular community), and from personal experience (whether the author's own or that of a witness).

j See <https://bbc.in/2FETOvU>

k See <https://bbc.in/3mqsp89>

l See <https://bbc.in/2H9u1wH>

m See <http://arg.tech>

n See <https://bbc.in/33CPFkm>

Each type of reasoning is associated with a specific template of critical questions pointing at the evaluation criteria for the reasoning.<sup>11</sup> In answering these questions, the user builds a confidence level in the support for the claim. Identifying any counter-considerations is another essential step in judging the impartiality of news articles. Again, the software helps students identify any such objections, often linguistically marked with indicative phrases such as “on the other hand,” “admittedly,” or “to some extent.”

In addition to a choice of five articles from various news sources across the political spectrum that are manually pre-analyzed by a team of experts to identify claims, reasons, and objections, The Evidence Toolkit employs automated methods for argument mining (also called argumentation mining in the literature) to allow the students to select their choice of article from the BBC News archives. The implemented argument mining technology automatically extracts the argumentative content from the news articles—provided the chosen article has any explicit argumentative content in it to begin with. Argument mining builds on the successes of Opinion Mining and Sentiment Analysis<sup>7</sup> to identify not only what views are being expressed in a text, but also why those views are held<sup>4</sup>—the software automatically processes the natural language text to produce the analysis otherwise performed by human experts.

At the time of this writing, The Evidence Toolkit has accumulated over 22,000 tries. The software has been well received, earning a rating of 4.15 out of 5 (where the average for applications on BBC Taster lies around 3.5). The user feedback moreover showed not only an accessible user experience (78% found it easy to use), but a successful one: 84% said the critical thinking tools explained in The Evidence Toolkit help to check the reliability of news, with 75% saying that it made them think more deeply about the topics at issue in the news articles. Putting critical literacy high on the BBC's agenda and applying argument technology to drive it also appears to reflect positively on the organization itself, with 73% stating that The Evidence Toolkit positively changed their view of the BBC.

The suite of argument technology developed for the BBC aims to address

the major societal challenge posed by intentional obfuscation and misinformation in the modern media landscape. In collaboration with the BBC—and with the producers of BBC Radio 4's Moral Maze in particular—we have approached critical literacy from several angles, developing quantitative debate analytics, interactive ways of engaging with argumentative material, and argument mining technology. With a distribution to over 3,000 educational institutions in the U.K., The Evidence Toolkit constitutes, to the best of our knowledge, the first public deployment of argument mining technology at scale. The further development of argument technology for reason-checking could provide a much needed weapon in combating fake news and reinforcing reasonable social discourse. **□**

#### References

1. Bex, F. et al. Implementing the Argument Web. *Commun. ACM* 56, 10 (Oct. 2013), 66–73.
2. Flaxman, S., Goel, S., and Rao, J.M. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly* 80 (2016), 298–320.
3. Johnson, R.H. and Blair, J.A. *Logical Self-Defense*. McGraw-Hill Ryerson, 1977.
4. Lawrence, J. and Reed, C. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.
5. Lazer, D.M.J. et al. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
6. Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220.
7. Pang, B. and Lee, L. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*. Now Publishers, 2008.
8. Sethi, R.J., Rangaraju, R., and Shurts, B. Fact checking misinformation using recommendations from emotional pedagogical agents. In A.Coy, Y. Hayashi, and M. Chang, Eds. *Intelligent Tutoring Systems 99*, 104 (2019), 99–104.
9. van Eemeren, F.H. et al. *Handbook of Argumentation Theory*. Springer, Cham, 2014.
10. Vosoughi, S., Roy, D. and Aral, S. The spread of true and false news online. *Science*, 359, 6380 (2018), 1146–1151.
11. Walton, D., Reed, C., and Macagno, F. *Argumentation Schemes*. Cambridge University Press, 2008.

**Jacky Visser** (j.visser@dundee.ac.uk) is a Lecturer in Computing with the Centre for Argument Technology, at the University of Dundee, in the U.K.

**John Lawrence** (j.lawrence@dundee.ac.uk) is a Lecturer in Computing with the Centre for Argument Technology, at the University of Dundee, in the U.K.

**Chris Reed** (c.a.reed@dundee.ac.uk) is Chair of Computer Science and Philosophy with the Centre for Argument Technology, at the University of Dundee, in the U.K.

This research was supported in part by EPSRC in the U.K. under grant EP/N014871/1, and in part by funding from the BBC. The authors would like to thank the entire ARGtech team that supported the development of software and resources for The Evidence Toolkit, and to recognize the input and support from Sharon Stokes, Head of BBC Young Reporter. The authors also want to acknowledge their enormous debt to Christine Morgan, Head of Radio, Religion and Ethics at the BBC, and to her production team on Radio 4's Moral Maze for their long-term support and enthusiasm for this initiative.

Copyright held by authors.





**ACM BOOKS**  
Collection II

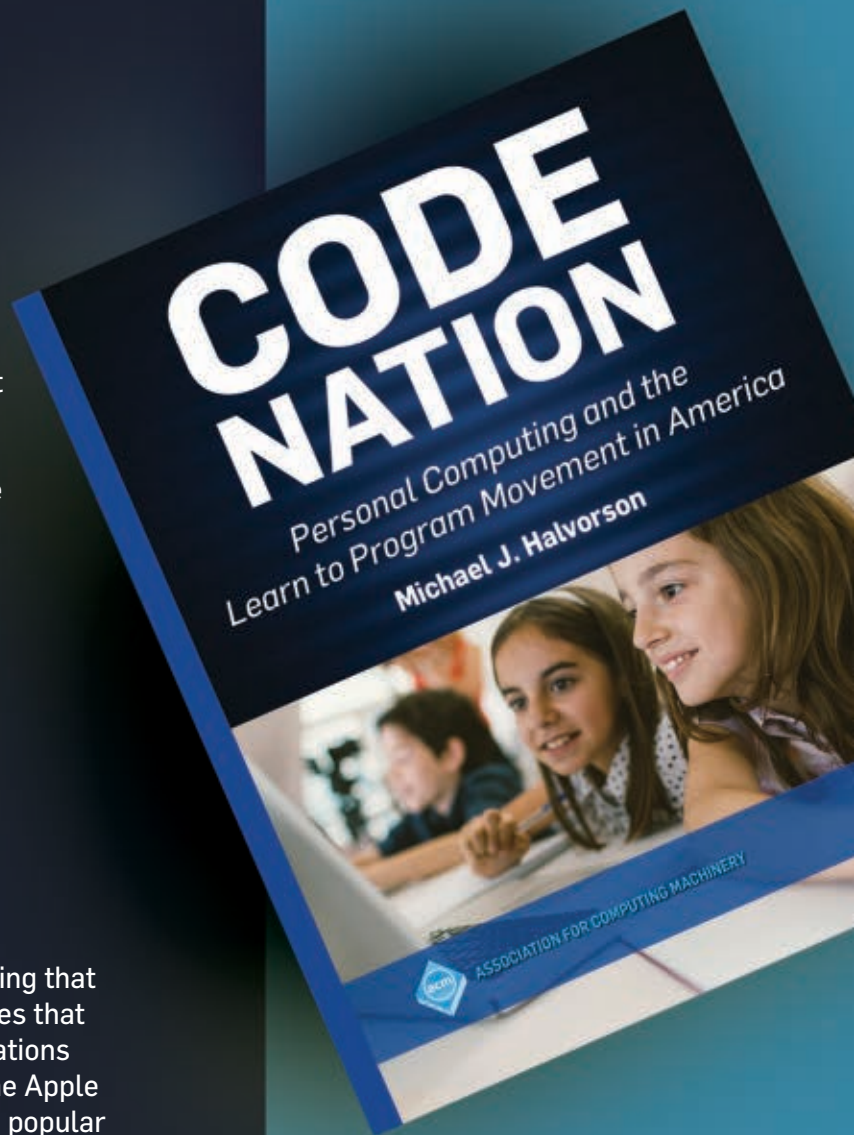
*Code Nation* explores the rise of software development as a social, cultural, and technical phenomenon in American history. The movement germinated in government and university labs during the 1950s, gained momentum through corporate and counterculture experiments in the 1960s and 1970s, and became a broad-based computer literacy movement in the 1980s. As personal computing came to the fore, learning to program was transformed by a groundswell of popular enthusiasm, exciting new platforms, and an array of commercial practices that have been further amplified by distributed computing and the Internet. The resulting society can be depicted as a “Code Nation”—a globally-connected world that is saturated with computer technology and enchanted by software and its creation.

*Code Nation* is a new history of personal computing that emphasizes the technical and business challenges that software developers faced when building applications for CP/M, MS-DOS, UNIX, Microsoft Windows, the Apple Macintosh, and other emerging platforms. It is a popular history of computing that explores the experiences of novice computer users, tinkerers, hackers, and power users, as well as the ideals and aspirations of leading computer scientists, engineers, educators, and entrepreneurs. Computer book and magazine publishers also played important, if overlooked, roles in the diffusion of new technical skills, and this book highlights their creative work and influence.

*Code Nation* offers a “behind-the-scenes” look at application and operating-system programming practices, the diversity of historic computer languages, the rise of user communities, early attempts to market PC software, and the origins of “enterprise” computing systems. Code samples and over 80 historic photographs support the text. The book concludes with an assessment of contemporary efforts to teach computational thinking to young people.

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



**CODE NATION**

*Personal Computing and  
the Learn to Program  
Movement in America*

**Michael J. Halvorson**

ISBN: 978-1-4503-7757-7

DOI: 10.1145/3368274



# Latin America Regional Special Section



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





# Welcome

**W**ELCOME TO THE special section on Latin America, covering all the Spanish- and Portuguese-speaking countries from Rio Grande to Cape Horn. Latin America is a striving region, with many countries aiming at a developed status in the near future, while at the same time facing enormous challenges in inequality, education, and government. The region also supports a great wealth of biodiversity. With generally less resources for research and a more difficult path for technology transfer when compared to developed countries, we aimed at highlighting the excellent level of research in computer science that flourishes in the region, both on basic research and on problems that are unique to Latin America.

We launched a general call for contributions welcoming research and development initiatives, large and small, aiming to cover as much as possible the diversity in development along the different countries. While countries like Argentina, Brazil, Chile, Mexico, and Uruguay exhibit, at different scales, promising environments for the development and advance of computing research, our survey also uncovered countries like Colombia, Costa Rica, or Peru, with a lot of potential that is waiting for stronger government or industry support in order to develop at large.

After a virtual workshop where an initial selection of the proposals was presented, we finally narrowed down the content to six “Big Trends” and eight “Hot Topics” contributions, with authors from Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Uruguay. We commissioned a journalist to complete the section by preparing a panoramic view of computer science in Central America and the Caribbean, with particular emphasis in Costa Rica. We thank all the colleagues that worked hard to present their research lines. While we had to leave out various interesting proposals, we believe the selection we present gives a good grasp of the landscape of computer science in Latin America, without being exhaustive.

The Big Trends articles cover the work of large research groups and/or labs with significant funding, on well-developed topics in the region like theoretical computer science, data management, supercomputing, dependable computing, digital health, and image processing. The Hot Topics articles describe initiatives of smaller groups, ranging from general research areas like natural language processing, smart cities, machine learning, data structures, and randomness services, to very special lines of research like estimating carbon stocks in Amazonia, characterizing Salsa music, or developing a system to fight the Coronavirus in one week.

We are confident readers will find the articles very exciting! We hope this section contributes to disseminate the wealth of activity in computer science research in a region where working on research is more difficult in many aspects, yet it still manages to stand out as a significant actor in the most important research areas of our discipline.

—*Virgilio Almeida, Gonzalo Navarro, and Sergio Rajsbaum*  
**Coordinators of Latin America Region Special Section**

**Virgilio Almeida** (virgilio@dcc.ufmg.br) is professor emeritus of computer science at the Universidade Federal de Minas Gerais, Brazil. He is also Faculty Associate at the Berkman Klein Center at Harvard University, Cambridge, MA, USA.

**Gonzalo Navarro** (gnavarro@dcc.uchile.cl) is a professor in the Department of Computer Science at the University of Chile, in Santiago.

**Sergio Rajsbaum** (sergio.ajsbaum@im.unam.mx) is a professor at the Instituto de Matemáticas at the Universidad Nacional Autónoma de México, in Mexico City, México.

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  
 Jakob Rehof  
 Haibo Chen  
 P J Narayama

### SPECIAL SECTION CO-ORGANIZERS

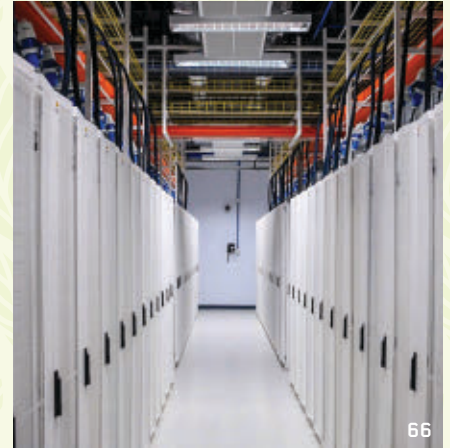
Virgilio Almeida  
 Universidade Federal de Minas Gerais and Harvard University  
 Gonzalo Navarro  
 University of Chile  
 Sergio Rajsbaum  
 Universidad Nacional Autónoma de México



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



## Hot Topics



- 46 **Estimating Amazon Carbon Stock Using AI-based Remote Sensing**  
*By Rosiane de Freitas, João M. B. Cavalcanti, Sergio Cleger, Niro Higuchi, Carlos Henrique Celes, and Adriano Lima*

---

- 49 **Why Me? Shedding Light on Random Processes via Randomness Beacons**  
*By Alejandro Hevia and Camilo Gómez*

---

- 51 **Toward Smart and Sustainable Cities**  
*By Fabio Kon, Kelly Braghetto, Eduardo Z. Santana, Roberto Speicys, and Jorge Guerra Guerra*

- 53 **A Technological and Innovative Approach to COVID-19 in Uruguay**  
*By Gastón Milano, Diego Vallespir, and Alfredo Viola*

---

- 56 **Contextualized Interpretable Machine Learning for Medical Diagnosis**  
*By Wagner Meira Jr., Antonio L.P. Ribeiro, Derick M. Oliveira, and Antonio H. Ribeiro*

---

- 59 **Understanding Salsa: How Computing Is Defining Latin Music**  
*By Carlos Arce-Lopera and Gerardo M. Sarria M.*

---

- 61 **Minding the AI Gap in LATAM**  
*By Barbara Poblete and Jorge Pérez*

---

- 64 **Three Success Stories About Compact Data Structures**  
*By Diego Arroyuelo, José Fuentes-Sepúlveda, and Diego Seco*

## Big Trends

- 66 **The Latin American Supercomputing Ecosystem for Science**  
*By Isidoro Gitler, Antônio Tadeu A. Gomes, and Sergio Nesmachnow*

---

- 72 **Digital Healthcare in Latin America: The Case of Brazil and Mexico**  
*By Monica Tentori, Artur Ziviani, Débora C. Muchaluat-Saade, and Jesus Favela*

---

- 78 **Chile's New Interdisciplinary Institute for Foundational Research on Data**  
*By Marcelo Arenas and Pablo Barceló*

---

- 84 **A Panorama of Computing in Central America and the Caribbean**  
*By Gerardo Torres Zelaya*

---

- 90 **Imaging Sciences R&D Laboratories in Argentina**  
*By Claudio Delrieux, Virginia Ballarín, Cristian García Bauza, and Mario A. López*

---

- 96 **A Tour of Dependable Computing Research in Latin America**  
*By Elias P. Duarte Jr., Raimundo Macêdo, Eliane Martins, and Sergio Rajsbaum*

---

- 102 **A Perspective on Theoretical Computer Science in Latin America**  
*By Marcos Kiwi, Yoshiharu Kohayakawa, Sergio Rajsbaum, Francisco Rodríguez-Henríquez, Jayme Luiz Szwarcfiter, and Alfredo Viola*



Environmental AI | DOI:10.1145/3416957

# Estimating Amazon Carbon Stock Using AI-based Remote Sensing

BY ROSIANE DE FREITAS, JOÃO M. B. CAVALCANTI, SERGIO CLEGER, NIRO HIGUCHI, CARLOS HENRIQUE CELES, AND ADRIANO LIMA

**F**ORESTS ARE THE major terrestrial ecosystem responsible for carbon sequestration and storage. The Amazon rainforest is the world's largest tropical rainforest encompassing up to 2,124,000 square miles, covering a large area in South America including nine countries. The majority of that area (69%) lies in Brazil. Thus, Amazonia holds about 20% of the total carbon contained in the world's

terrestrial vegetation.<sup>1,5,7</sup> But the rampant deforestation due to illegal logging, mining, cattle ranching, and soy plantation are examples of threats to the vast region. Biodiversity loss, ecosystem imbalance, and higher concentration of carbon dioxide in the atmosphere are related consequences.<sup>9</sup>

Based on the directives given by the Intergovernmental Panel on Climate Change (IPCC), there is an urgent need to provide additional guidance on the

design of forest monitoring systems. This involves issues such as forest inventory design, stratification, sampling, pools, accuracy/uncertainty assessment, and the combination of ground-based inventories with remote sensing and modeling approaches. Computing approaches can provide valuable tools to support the development of efficient solutions for this environmental problem. In this article, we present some ongoing research initiatives to address the carbon stock estimation problem.

We are interested in estimating carbon stocks by means of extrapolation and spatialization based on ground-based forest inventory combined with heterogeneous sources of remote sensing images through high-resolution satellite, radar, and LiDAR (Light Detecting and Ranging) 3D technology. One of our objectives consists of determin-

ing small areas of forest with high density of captured carbon through the application of artificial intelligence (AI) strategies involving pattern recognition, graph theory, image retrieval, machine learning, and combinatorial optimization techniques. In this article, we give some details of this work. In order to refine the carbon stock estimate, we present related research work addressing the detection of clearings in the Amazon rainforest based on satellite and radar images using machine learning techniques.

## Identification of Representative Plots for the Optimization of the Carbon Capture Estimation Process

The forest inventory is the collection of attributes about the quantitative and qualitative characteristics of the forest, providing information on forest resources that are applied to monitoring,

**This work presents advances in the way of a more accurate estimate of the carbon captured by forest areas, in particular the Amazon rainforest and its peculiarities.**



forest management policies, as well as strategic actions to exploit resources in a sustainable way. Along with this information, we highlight the measures of biodiversity, social aspects of the forest and biomass, and carbon stocks. On the other hand, the biomass of a forest is the quantity, by mass, of living or dead matter, present in the vegetation or only in its arboreal fraction. In general, it is measured by allometric equations developed by forest engineers. These equations make use of forest attributes given as inputs for their calculation, such as tree height, vegetation index, diameter at breast height (DBH) and tree crown diameter, among others (Figure 1). Its estimation is useful and stands out as an ecosystem assessment tool. With it, it is possible to carry out analyzes of productivity, energy conversion, nutrient cycling, absorption and storage of solar energy, as well as estimating the carbon storage. The latter can be estimated from forest biomass because the forest absorbs and stores carbon in its mass when it is in its development and growth phase. Therefore, forests act as sinks of carbon when they are in their phase of expansion and development. We developed a mobile application for the automatic determination of DBH (three-dimensional structure) of a tree, through the analysis of the photo from the camera of a smartphone applying image retrieval/processing and computational geometry techniques.

Thus, given a forest region to be inventoried to estimate the carbon stock, it is important that the samples, or plots, be installed in places where there are more representative trees:

the dominant (for example, the largest) and the emerging (tallest) trees. In the field survey, only the dominant trees are identified. On the other hand, using data from LiDAR images (Figure 2), it is possible to identify the emerging ones.<sup>4,6</sup> In a more representative way, we can consider small areas with a higher density of captured carbon based on the average height of trees in a forest fragment. For this, it is necessary to determine the Digital Surface Model (DSM) and Digital Terrain Model (DTM) of each forest fragment (Figure 3).

Thus, the problem of determining the most representative forest fragments was modeled as a Maximal Covering Location Problem (MCLP), which is NP-hard.<sup>2,3</sup> In this problem, we have a set of plots from which the most representative ones will be extracted. Therefore, in the MCLP model the facilities will be the most representative plots and the demand will be each of the candidate plots. The representativeness index of the plots was calculated based on the sum of the average height of the trees of its  $n$  adjacent plots. A hybrid strategy combining an IP formulation and Greedy Randomized Adaptive Search Procedure (GRASP)<sup>10</sup> indicates potential large-scale use.

Based on the amount of captured carbon calculated in the permanent sample plots forestry periodically inventoried, it was possible to carry out a comparative analysis with the estimated through the MCLP on the LiDAR point clouds of such regions (Figure 4), indicating the suitability of the strategy adopted. Extrapolations, considering the total Amazon rainforest area, indicates



Figure 1. DBH measure of a tree in hard conditions.



Figure 2. 3D high-resolution remote sensing by LIDAR.



Figure 3. DSM and DTM calculate the average height of a segment of forest.

an increase of at least 4% to 10% in the total amount of Amazon carbon stock, which would give up to 30% of the total captured in terrestrial vegetation areas around the world.

#### Automated Detection of Deforestation Areas in the Amazon Region Using Remote Sensing and Machine Learning

This line of research addresses the problem of automated detection of

deforestation areas in the Amazon rainforest. Currently, there are image classification systems in use for detecting deforestation areas. However, small-scale clearing is a challenge that hinders detection from satellite monitoring. We have proposed an approach for classifying remote sensing images, comprising three steps: image segmentation, feature extraction, and classification.

Different techniques can



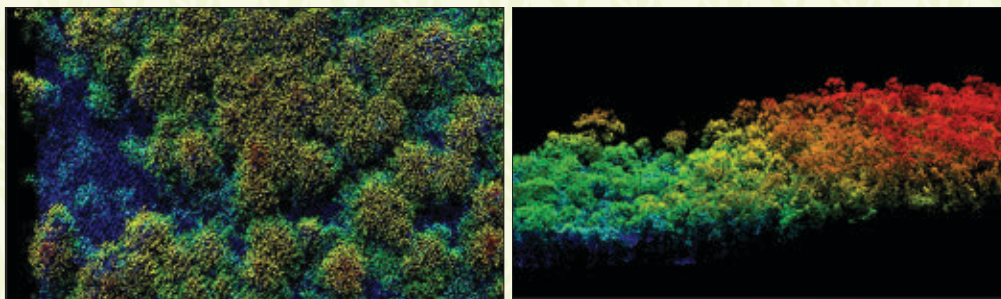


Figure 4. LIDAR point cloud of a segment of the Amazon rainforest.

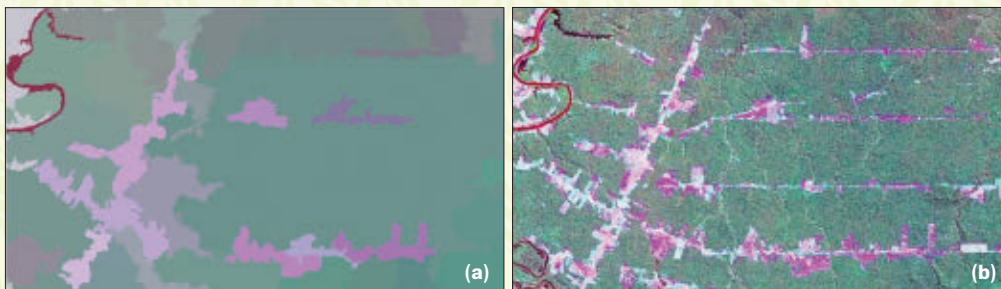


Figure 5. A satellite image example and its corresponding segmentation applying (a) supervised and (b) unsupervised learning techniques.

be used in each of these steps. Experiments were performed with several classification algorithms, seven supervised techniques: SVM, decision tree, perceptron, random forest, logistic regression, KNN and naive Bayes; and two unsupervised ones:  $k$ -means and BIRCH. The goal is to find a classification model that best describes distinct patterns of deforestation in the images.

We have worked with a set of satellite images (Landsat™) characterized by color and texture measurements that constitute the features used for classification. The experiments performed used images from Barcelos (a municipality near the city of Manaus, the capital of the state of Amazonas in Brazil), comprising 26 images divided in 3.288 segments with at least 700 pixels each. The corresponding ground truth dataset is made of 1,636 clearing and 1,652 no-clearing segments. Figure 5 presents an example of an image segmentation. Note the segmentation approxi-

mately separates the areas with deforestation. This facilitates the classification step. There are 781 features per image segment.

We obtained encouraging results from the experiments. Among the seven supervised classification techniques tested, the decision tree technique reached the highest accuracies of 97.18 and 97.65 for deforestation and no-deforestation segments, respectively. Among the unsupervised approaches ( $k$ -means and BIRCH) both reached an accuracy of approximately 95% also for both deforestation and no-deforestation segments.<sup>8</sup>

Note, however, that this is a preliminary study with a small dataset. Nevertheless, given the vast majority of the Amazon region presents similar visual features, the results presented here are very promising.


### Conclusion

We present two approaches that allow the refinement of the estimated carbon captured by the forest. The

traditional techniques for estimating carbon are very general and are based on large areas of vegetation cover not taking into account discontinuities, such as clearing areas or the survey of small areas with a higher concentration of carbon. Thus, this work presents advances in the way of a more accurate estimate of the carbon captured by forest areas, in particular the Amazon rainforest and its peculiarities.

We have also developed an automated technique for the detection of deforestation areas in the Amazon. Although the image base tested is small, the results obtained are encouraging. It is necessary to perform experiments with a larger number of satellite and radar images from different areas in order to draw conclusive results. This can also help the carbon stock estimation based on the variation of deforestation in certain areas.

The goal of this article was to illustrate how the application of computa-

tional strategies based on AI, combined with high-resolution remote sensing, can provide valuable tools for efforts in environmental conservation. 

### References

1. Brienen, R. et al. Long-term decline of the Amazon carbon sink. *Nature* 519 (2015), London, 344–348.
2. Church, R. And Reville C. The maximal covering location problem. *Papers of the Regional Science Association* 32 (1974), 101–118.
3. Farahani, R. et al. Covering problems in facility location: A review. *Computer and Industrial Engineering* 62, 1 (2012), 368–407.
4. Freitas, R. De, Silveira, D. and Higuchi, N. Estimating the carbon stocks by optimizing LiDAR forest big data. *eScience Workshop*, 2016.
5. Gaul, T. et al. Long-term effect of selective logging on floristic composition: A 25-year experiment in the Brazilian Amazon. *Forest Ecology and Management* 440 (2019), 258–266.
6. Junttila, V. et al. Strategies for minimizing sample size for use in airborne LiDAR-based forest inventory. *Forest Ecology and Management* 292 (2013), 75–85.
7. Malhi, Y. et al. The regional variation of aboveground live biomass in old-growth Amazonian forests. *Global Change Biology* 12 (2006), 1–32.
8. Pessoa, M.; Cleger, S., Cavalcanti, J. M., and Freitas, R. De. Detecção de áreas de clareiras na Floresta Amazônica através de monitoramento via satélite usando técnicas de aprendizagem de máquina. In *XLVI Seminário Integrado de Software e Hardware*. Belém-PA. Anais do XLVI Seminário Integrado de Software e Hardware, Congresso da Sociedade Brasileira de Computação. Porto Alegre (2019) 125–136.
9. Peterson, C. et al. Critical wind speeds suggest wind could be an important disturbance agent in Amazonian forests. *Forestry* (2019), 1–16.
10. Resende, M., Ribeiro, C. *Optimization by GRASP: Greedy Randomized Adaptive Search Procedures*. Springer-Verlag, New York, 2016.

**Rosiane de Freitas** is an associate professor at PPGI and Instituto de Computação, Universidade Federal do Amazonas, Manaus, Amazonas, Brazil.

**João M.B. Cavalcanti** is an associate professor at PPGI and Instituto de Computação, Universidade Federal do Amazonas, Manaus, Amazonas, Brazil.

**Sergio Cleger** is a researcher at PPGI, Manaus, Amazonas, Brazil.

**Niro Higuchi** is Principal Research Coordinator at Laboratório de Dinâmica e Manejo Florestal, Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brazil.

**Carlos Henrique Celes** is a researcher at Laboratório de Dinâmica e Manejo Florestal, Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brazil.

**Adriano Lima** is a researcher at Laboratório de Dinâmica e Manejo Florestal, Instituto Nacional de Pesquisas da Amazônia, Manaus, Amazonas, Brazil.



# Why Me? Shedding Light on Random Processes via Randomness Beacons

BY ALEJANDRO HEVIA AND CAMILO GÓMEZ

**W**E LIVE SURROUNDED by random processes, systems whose final outcome are typically unpredictable. When drawing a raffle, playing roulette at the casino, deciding who pays the restaurant bill, or even gambling in a poker game, we take part of a random process. Indeed, their intrinsic unpredictability is often what drives us to participate in them.

There are situations, however, where the result of these random processes may have serious consequences for those involved. Being selected for a tax audit, for example, can carry significant time and legal costs, which is why the selection process must be trustworthy. We accept a tax audit if we believe the random process was not manipulated in any way, that their final outcome was not influenced in an

improper manner. But, if the results are potentially unpredictable, *any* result may be likely. How do we then prove that some specific outcome was not deliberately chosen? Solving this apparent paradox is the objective of a verifiable randomness service offered by the University of Chile.

Funded by a grant from the U.S. Department of Commerce and based on a proposal by the National Institute of Standards and Technology (NIST), Random UChile<sup>a</sup> provides verifiable randomness through a once-a-minute online pulse, a service that generates one 512-bit value every 60 seconds. This public random value is not only generated in a robust, unpredictable, and consistent way, but also the execution of the entire process is verifiable. The system, also known as a *randomness beacon*,

<sup>a</sup> <https://random.uchile.cl/en>

has been designed to be transparent and open—the most recent value can always be verified by any external observer, yet the next output value remains unpredictable. This correct randomness generation is possible thanks to a variant of a cryptographic algorithm design by NIST.<sup>2</sup> The pulses generated by this algorithm are then added to a signed hash chain, preventing malicious manipulation of previously posted values.

The system relies on polling random data continuously from several entropy sources, both internal (TRNG hardware) and external (online) such as the Centro Sismológico Nacional (National Seismological Center) of Chile, which indicates the characteristics of Chilean earthquakes; the University of Chile's streaming radio, which transmits live shows and music; Twitter, which supplies a stream of randomly selected tweets in real time; and the Ethereum blockchain in the form of the hashed value of its last block. One can see the beacon as turning unpredictability (earthquake, tweets, or any of the others) into a guarantee of randomness of the output values. Indeed, some key innovations and cryptographic tools are needed for this to happen.

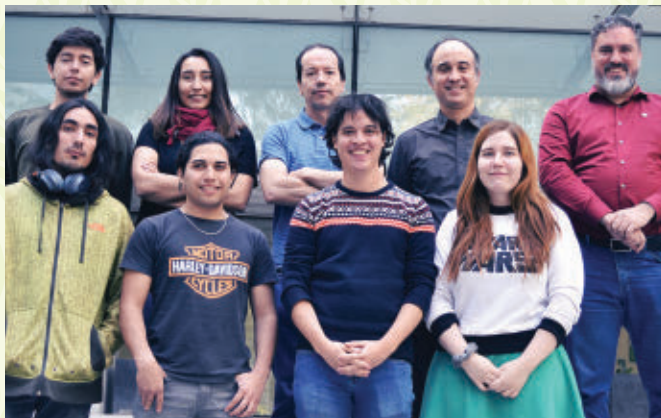
To foster the creation of new beacons, both the technical design of the Random UChile's beacon and the cryptographic analysis that supports the claim of verifiability, are being shared with the research community.<sup>1</sup>

## Improving Transparency in Public Organizations

One key use case of verifiable randomness is ensuring the correctness of those processes run by government agencies, particularly those that rely on randomized algorithms for their decisions. Toward this end, Random UChile has collaborated with the Comptroller General (or Contraloría General de la República, CGR) in Chile, the public agency in charge of auditing the tax expenditure in the country. The CGR must periodically select, at random, who among authorities and public servants must be subject to an audit. Without verifiable randomness, the CGR was exposed to accusations of political persecution. In response, the CGR developed a pilot program to use Random UChile's verifiable randomness, where fair selection follows from the guarantees behind the beacon design.

Random UChile can also be used to actively prevent

**Random UChile provides verifiable randomness through a once-a-minute online pulse, a service that generates one 512-bit value every 60 seconds.**



**The Random UChile group. Top row, from left: Juan Rojas, Constanza Csori, Sergio Miranda, Alejandro Hevia, and Cristián Rojas. Bottom row, from left: Franco Pino, Alejandro González, Camilo Gómez, and María José Vilches.**

conflicts of interest that otherwise may encourage litigation or, worse, increase distrust on the underlying public processes. One such example is the use of Random UChile’s verifiable randomness to ensure fair selection of judges for legal disputes. A pilot of the system is being considered for domain name registration controversies at the .CL administrator.

### Impact on Elections Systems

The electoral voting system in Chile relies on the work of *vocales*, poll workers that are regular citizens chosen to perform as official workers during the days of the election. Unfortunately, in every election there are well-publicized controversies about the selection of poll workers. Even though they are supposed to be selected at random among all able citizens, often poll workers complain they have been selected for three, or more, elections in a row (although the law may allow such cases to occur under limited circumstances). The opacity of this process hurts democracy as partisan selection of these roles may open the door for corruption, or at least distrust on the

election outcome. With the help of verifiable randomness, we can improve the transparency of the poll worker selection process and provide a way for citizens to confirm that every nomination was indeed fair and correct.

Another use of verifiable randomness comes from auditing elections. Recent concerns about the security and trustworthiness of electronic voting systems used in the U.S. and other countries have prompted the development of post-election audits. Risk-limiting audits<sup>5</sup> is a well-known technique to verify an election after it has completed without doing a total recount of the votes, only counting a much smaller random sample. The use of verifiable randomness not only makes it more efficient (no need of old-fashion dice throwing) but also more transparent at a larger scale-verification results can now be inspected online by interested citizens and organizations.

### Using Randomness Beacons in Complex Systems


Verifiable randomness is difficult to obtain but extremely useful in

the design of fair and transparent systems. Besides allowing transparency in new applications (for example, providing fully auditable statistical sampling for verifiable scientific experiments), verifiable public randomness is likely to become a public utility, a service upon which new and more sophisticated protocols and systems will be built. Examples include faster and more secure crypto-currencies,<sup>3</sup> lottery systems based on verifiable yet private randomness, and privacy-preserving verifiable data management systems based on secure multiparty computation.<sup>4</sup> Random UChile is contributing to the effort by actively developing prototypes for some of these systems.

### Similar Projects

NIST’s Interoperable Randomness Beacons project<sup>b</sup> has become one crucial driver in the creation of new beacons. It not only maintains a beacon implementation<sup>c</sup> and provides the official guidelines for implementing interoperable randomness beacons, but also explicitly seeks to “promote the deployment of Beacons by multiple independent organizations.” Luckily, NIST’s efforts seem to be paying off. Following the Random UChile project, the Inmetro Randomness Beacon<sup>d</sup> in Brazil has joined the cause of reliable public randomness. Since all three beacons follow the NIST reference,<sup>2</sup> applications built on top of any of these services have now plenty of choices to deposit their trust.

b <https://csrc.nist.gov/projects/interoperable-randomness-beacons>  
 c <https://beacon.nist.gov/home>  
 d <https://beacon.inmetro.gov.br/>

Yet trust is hard to gain. To achieve trustworthy public randomness without trusting on a single entity, Random UChile is also contributing to a global project called The League of Entropy,<sup>e</sup> a large-scale effort by several organizations across the world that seeks to produce a distributed randomness beacon. In this system, the public randomness is produced by the Drand Protocol,<sup>6</sup> which combines the output of all participants. The result is guaranteed correct and fair as long as at least one entity follows the rules and correctly contributes to the randomness. From local trust to global trust, the journey is just starting. 

e <https://www.cloudflare.com/leagueofentropy/>

### References

1. Gómez, C., Hevia, A., Miranda, S., Riveros, E., and Rojas, C. Design and Implementation of a Verifiable Public Randomness Beacon. Technical Report, submitted 2020.
2. Kelsey, J., Brandão, L. T.A.N., Peralta, R. and Booth, H. A Reference for Randomness Beacons: Format and Protocol Version 2. Draft NISTIR 8213 Publication. May 2019. <https://csrc.nist.gov/publications/detail/nistir/8213/draft>.
3. Kiayias, A., Russell, A., David, B. and Oliynykov, R. Ouroboros: A provably secure proof-of-stake blockchain protocol. In *Proceedings of the Annual Intern. Cryptology Conf.* (2017). Springer, 357–388.
4. Lindell, Y. Secure Multiparty Computation (MPC). IACR Technical report 2020/300; <https://eprint.iacr.org/2020/300>
5. Norden, L., Burstein, A., Hall, L.J. and Chen, M. Post-election audits: Restoring trust in elections. Technical report. Brennan Center for Justice and Samuelson Law, Technology & Public Policy Clinic, 2007.
6. Syta, E., Jovanovic, P., Kogias, E.K., Gailly, N., Gasser, L., Khoffi, I., Fischer, M.J. and Ford, B. Scalable bias-resistant distributed randomness. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy* (San Jose, CA, 2017). IEEE Press, 444–460.

**Alejandro Hevia** is an assistant professor in the Department of Computer Science at the University of Chile, Santiago.

**Camilo Gómez** is Random UChile Coordinator in the Department of Computer Science at the University of Chile, Santiago.



# Toward Smart and Sustainable Cities

BY FABIO KON, KELLY BRAGHETTO, EDUARDO Z. SANTANA, ROBERTO SPEICYS, AND JORGE GUERRA GUERRA

**L**ATIN AMERICA HOSTS some of the world's great metropolises, with a plethora of social problems facing the complex societies that live there. Not only is there a lack of proper infrastructure but also there is a high degree of inefficiencies in urban services and a lack of effective management. Evidence-based public policymaking is finally starting to gain attention as governments and academic projects throughout the region begin to apply modern computer science techniques to develop tools for both operating the city's daily life and guiding long-term management. Instead of focusing on futuristic smart cities built from scratch or on improving user experience in rich cities, Latin American researchers pay

attention to underprivileged neighborhoods and their low-income populations, leveraging existing data and collecting new datasets to support better decision-making.

A prime example is the InterSCity project,<sup>a</sup> a consortium of 11 universities and startups in Brazil in which computer scientists work together with architects, urban planners, transportation engineers, economists, and health professionals. The goal is to produce innovative science and open source software tools to address relevant urban problems using Internet of Things (IoT) technologies, high-performance computing, big data analytics, and visualization.

Based on the study of tens of projects around the world, InterSCity research-

<sup>a</sup> <https://interscity.org>



ers identified the most relevant requirements for software platforms for the development of smart city applications.<sup>3</sup> They proposed a reference architecture containing the basic elements that should be provided to enable the rapid development of applications for citizens, governments, and urban service providers. An open source implementation of this architecture has been developed with scalability as its major goal, to be able to handle millions of user simultaneously.<sup>1</sup> Testing such large-scale deployments in real life is usually impractical, thus the project also developed a simulator capable of simulating over 10 million

agents acting in a city.<sup>b</sup>

These tools have been used both to educate the next generation of developers and scientists in universities and to create prototypes and pilot studies. Such a focus on education is also perceived in Peru, where the IoT Research Group at the National University of San Marcos<sup>c</sup> has implemented changes in the CS curriculum to incorporate IoT and smart cities topics in basic CS education.

Startup companies founded by alumni from such research groups can be an effective means to intro-

<sup>b</sup> <https://interscity.org/software/intersimulator>

<sup>c</sup> <http://iotunmsm.pide.gob.pe>

**Latin American researchers pay attention to underprivileged neighborhoods and their low-income populations, leveraging existing data and collecting new datasets to support better decision-making.**



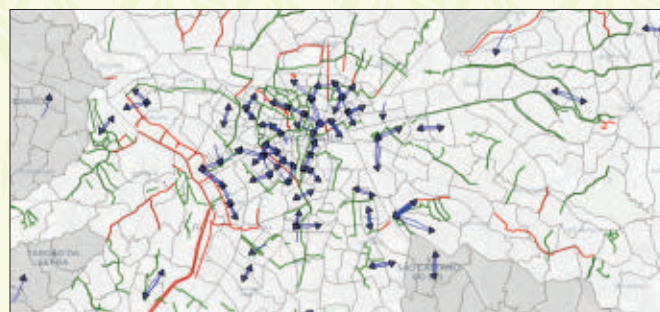
duce these ideas into society. The Scipopulis startup, for example, developed a citizen app (Figure 1) using advanced machine learning and HPC techniques to provide accurate estimates of bus arrival times for each of the 20,000 bus stops in São Paulo every minute of every day. The technology was extended and implemented as a Real-Time Bus Dashboard system (Figure 1) that was deployed in the city to help manage its 15,000 bus fleet, one of the largest in the world, which serves nearly nine million passengers daily. Similar efforts are underway in the fields of public health, accessibility, and cycling.

Urban cycling offers a healthy, environmentally friendly means of transportation that could easily serve from 10% to 20% of the trips in a city. In partnership with the São Paulo City Traffic Engineering Company, InterSCity developed an open source tool<sup>d</sup> capable of analyzing data from millions of bike trips, gathered from bike-sharing systems and mobility surveys. The tool produces rich visualizations of mobility flows against existing cycling infrastructure (Figure 2) that provide insights for city planning.

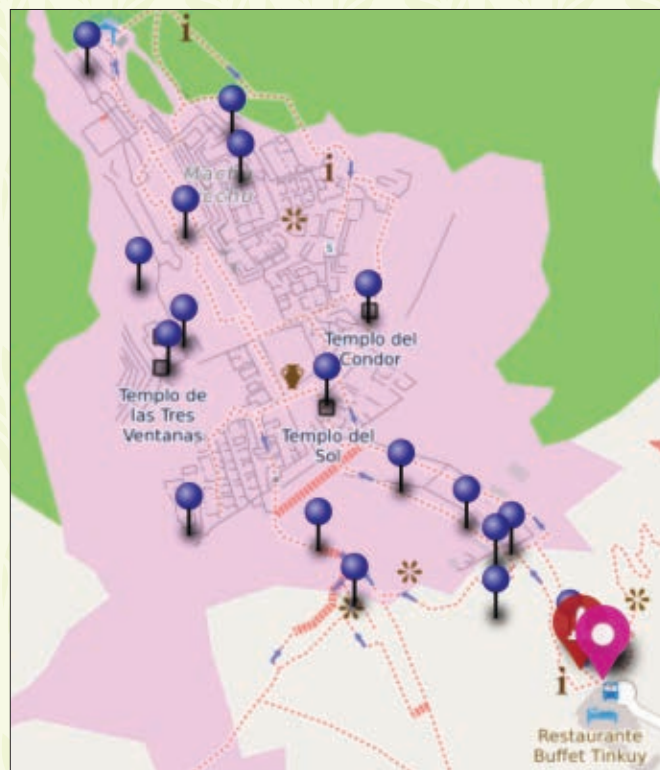
Green urban areas are essential for providing good quality of life in cities. Unfortunately, they are a resource that is often scarce, mainly in underprivileged zones. Finding water for the irrigation of these areas might also be a problem as good quality water supply is often limited. In Lima, Peru, researchers from the National University of San Marcos are deploying an IoT system that uses open



**Figure 1. (left) ML-based citizen app. (right) Real-Time Bus Dashboard for system operators.**



**Figure 2. Bike science tool displaying cycling mobility flows.**



**Figure 3. Proposed Beacon network in Machu Picchu.**


hardware with humidity and temperature sensors to monitor public parks and determine when and how much water to dispense in the green areas.<sup>2</sup> The idea is to provide adequate irrigation while minimizing the use of water.

In the historic Inca city of Machu Picchu, researchers are seeking resources to deploy a network of beacons to monitor in real-time the movements of tourists across the city (Figure 3). The collected data will help provide a better

understanding of tourists' movements throughout the archeological area and promote a better experience while taking care to preserve and maintain these historical sites.

Computer scientists have an unprecedented opportunity to work toward improving the quality of life of billions of people in contemporary cities. Computational tools can provide innovative services for citizens and elements to support evidence-based public policy-making for governments. To accomplish that, it is fundamental to promote high-quality R&D as well as make politicians and the overall population aware of the importance of using science as the base for effective governance and social and economic development.

**Acknowledgments**

InterSCity is funded by CNPq 465446/2014-0, CAPES, and FAPESP procs. 14/50937-1 and 15/24485-9. 

**References**

1. Esposte, A. et al. Design and evaluation of a scalable smart city software platform with large-scale simulations. *Future Generation Computer Systems* 93 (Apr. 2019), 427–441.
2. Guerra, J. et al. Implementación de un servicio de monitoreo y control de jardines en una municipalidad de Lima usando Internet de las Cosas. *Congreso Internacional de Ingeniería de Sistemas*. Peru, 2019.
3. Santana, E. et al. Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. *ACM Computing Surveys* 50, 6 (Nov. 2017), Article 78.

**Fabio Kon** is a professor at the University of São Paulo, Brazil.

**Kelly Braghetto** is an assistant professor at the University of São Paulo, Brazil.

**Eduardo Z. Santana** is a research scientist at the University of São Paulo, Brazil.

**Roberto Speicys** is Chief Scientist at Scipopulis Ltd. in São Paulo, Brazil.

**Jorge Guerra Guerra** is a professor of computer science at National University of San Marcos, Lima, Peru.

<sup>d</sup> <https://gitlab.com/intercity/bike-science>



# A Technological and Innovative Approach to COVID-19 in Uruguay

BY GASTÓN MILANO, DIEGO VALLESPER, AND ALFREDO VIOLA

**T**HIS ARTICLE PRESENTS a technological and innovative approach developed to help the Uruguayan government in their fight against COVID-19. The first version of the system (with only the most urgent services at that time) was released only seven days after the first case of COVID-19 was reported in Uruguay! At press time, some months after its first release, the fourth version is operative. Part of the system is a cellphone app available freely to the public, and it was downloaded by half a million people in a country with a population of 3.5 millions.

The project is innovative because it is the only worldwide solution, that we are aware of, that integrates in a unified way for patients, all health services of a country, the Ministry of Health, self-monitoring, remote patient monitoring, and telemedicine. Furthermore, the system makes full tracking possible from end to end to follow citizens' and patients' situation. Because of this, Uruguay was one of



the first three countries and the first in Latin America to incorporate exposure notifications for COVID-19.

As we write this article (July 2020), the world is immersed in a context of total uncertainty regarding health issues caused by the coronavirus. In Uruguay, the first case of COVID-19 was confirmed on March 13, 2020. The same day, a CoronaVirus UY Plan was launched by the Uruguayan government, where technol-

ogy (as in South Korea and other countries) had to play a key and strategic role.

Some of the main objectives to be achieved were:

- ▶ avoid the collapse of the health system;
- ▶ avoid having to directly contact the possibly infected people by phone calls;
- ▶ be proactive in managing the epidemic;
- ▶ the software products had to be robust and extremely secure, due to the data they would handle; and,
- ▶ have an operative solution as soon as possible.

To achieve these objectives, a team of private and public companies collaborated as well as interested individuals, organizing themselves to carry out a software engineering pro-

ject never seen before, at least in Uruguay.

The Uruguayan government (through the Presidency Board) asked the group to work quickly to achieve a first delivery in less than seven days. This first version had to include at least registration, classification based on epidemiology surveys that would make it possible to understand which citizens were more likely to have the virus, and it had to communicate about them to different health providers (using multiple channels) by telephone or a video call protecting in this way the health personnel.

This was, as the reader has probably already noticed, a mission-critical project to be executed in a few days.

**Uruguay is one of the first countries to incorporate contact tracing for COVID-19.**



## The Coronavirus.uy System

In order to avoid the collapse of the communication channels in the health care systems, a multi-channel solution (for receiving automated evaluation from the citizens or a request for information about COVID-19) was developed. During these seven days, the team built native mobile apps for Android<sup>a</sup> and iOS,<sup>b</sup> a Progressive Web application,<sup>c</sup> a form to be used by call centers, chatbots integrated in several conversational channels: Whatsapp,<sup>d</sup> Facebook Messenger, and WebChat embedded in the government pages. The team also developed epidemiological risk calculation modules, a survey module, a registration module, and the implementation of workflow processes for the correct handling of data by health providers.

Figure 1 shows one of the several workflows of the system. The user sends her data and the system pro-

- a coronavirus.uy cell app at Google Play; <https://play.google.com/store/apps/details?id=uy.gub.salud.plancovid19uy>
- b coronavirus.uy cell app at Apple Store; <https://apps.apple.com/us/app/coronavirus-uy/id1503026854>
- c <https://servicios.coronavirus.gub.uy/app/home>
- d <https://wa.me/59898999999>

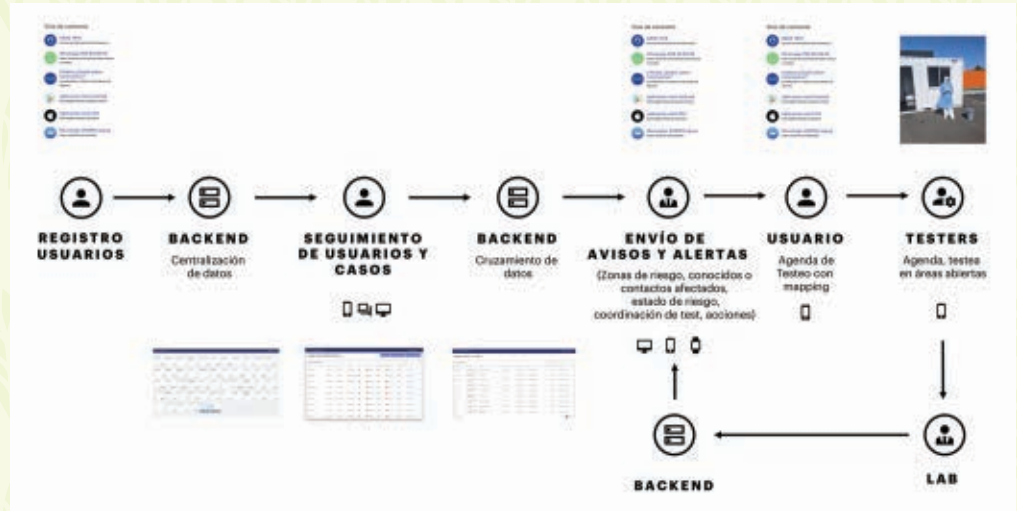


Figure 1. Healthcare provider request handling process.

cesses it in different levels, also providing the possibility of a drive through coronavirus test to be scheduled.

More than 30 companies (private and public) were involved in developing the system in the seven-day deadline set by the government. The team's roles were security managers, UX designers, healthcare workers and healthcare providers from the Ministry of Health and Epidemiologists, and around 40 people working as programmers, software architects, testers, and managers. The main programming language used was GeneXus<sup>e</sup> (a low code tool) generating code in Java, MySQL (SQL relational storage), Swift (iOS),

- e <https://www.genexus.com/en>

**It is important to emphasize that not only has Uruguay's health system *not* been saturated, but also 100% of COVID-19 suspected cases have been handled by coronavirus.uy.**

Kotlin, and Java for Android and Javascript.

On the first day, the team could not know whether this was going to result in a successful project or not. The majority of the people involved had never worked together before, no development process was established, there were no channels of communications between people or companies and, primarily, there were no written functional specifications. But it did, it worked! Figure 2 shows members of the development team in virtual meetings.

Preserving privacy and information security was a design decision from the beginning. The national e-government agency (AGESIC)<sup>f</sup> demanded not only these conditions but also specialized equipment in the project to ensure they were enforced.

Apple and Google have acknowledged that exposure notification was integrated in the last version of the coronavirus.uy app.<sup>g</sup> The first exposure notifications were sent at

the beginning of July 2020 and the people whom received them used the system to be in contact with the health care provider, who finally took the coronavirus tests.

One remarkable fact is that while in countries like Germany, where their exposure notification solution costs \$22.5 million,<sup>h</sup> the entire coronavirus.uy system (not only exposure notification!) has no cost at all. In Uruguay, the developers worked for free and the whole system (not only the phone app) belongs to the government.

This was a very difficult project with an extremely tight deadline. This achievement was possible thanks to the particular (and unique) Uruguayan ecosystem in software development. A solid education in software engineering, a highly digitized government, an integrated and digitally strong healthcare system, the use of a low-code platform (GeneXus) for a rapid development of the system, and other important advantages that the country offers, are at the basis of this "Uruguayan miracle."

- f <https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/>
- g <https://bit.ly/39AWNjL>

- h <https://bit.ly/2Db5ipR>





Figure 2. Members of the development team connected in virtual meetings.

## Results

Two months after the first case of COVID-19 was reported in Uruguay, the government and the population have an integrated (and evolving) system to handle this pandemic. Furthermore, the system is very flexible and can be set up to cover similar crises in the future.

From a total population of 3.5 million people (2.5 million active), the system has been accessed more than 2.6 million times (through the different communication channels offered). In other circumstances, the phone system of the country would have collapsed! The novel approach (in contrast with all other technological solutions in the world) is the fact that the citizen does not call the health provider if symptoms of COVID-19 are suspected. To the contrary, it uses the technological solution that automatically contacts the appropriate health provider. Then, based on their priorities (for example, risk factors or health symptoms recorded in the online form), the provider calls the citizen. It is important to emphasize that not only has Uruguay's

health system *not* been saturated, but also 100% of COVID-19 suspected cases have been handled by coronavirus.u.y.

The design and success of coronavirus.u.y was a key element for Google and Apple to contact Uruguay to incorporate their contact tracing software for Exposure Notification in it. Its fourth version includes a contact-tracing system in order to help, even more, the population and the Government in this pandemic crisis. Moreover, coronavirus.u.y has been requested by other countries like Brazil, Paraguay, Peru, and Surinam.

Regarding privacy, the Agency for Electronic Government in Uruguay (AGESIC) published all the code of this solution<sup>i</sup> to allow the code to be reviewed and inspected. Also, the team has been in contact with several universities to give early access to the code to allow for auditing and to ensure privacy preservation.

A key contribution of coronavirus.u.y to the public health system in Uruguay (that serves one third of the population) is the instal-

<sup>i</sup> <https://bit.ly/3g9cPUC>


lation of an inbox-driven data tracking tray to attend the suspected cases automatically. Moreover, these tools will be available in the future to be used in their general needs beyond the virus. In addition, the telemedicine integrated in the system, is protecting doctors and nurses by avoiding contact with infected people without life-threatening complications. This technological innovation improved the quality of attention paid to the country's public health system.

## Conclusion and Future Work

Uruguay has managed to have an essential software system in virus management in record time. Thirty companies worked together, with no written functional requirements, with no clear software engineering process in place, and also, without having in place other aspects that we normally consider essential in any software development project. In spite of these constraints, the system is working, it incorporates innovative worldwide features compared to other similar technological solutions, and it is evolving. Coronavirus.

uy is now an asset of the Uruguayan people and of the state. Furthermore, it can handle any kind of epidemic with a small configuration change in the system.

For future work we plan to analyze what has happened behind the curtains. We want to investigate which non-written processes were followed, how, and when the team communicated, how they handled the non-written requirements situation and how they managed and controlled this project. We expect to find a highly agile process with an extreme commitment of the developers in order to achieve their goals.

The world is clearly changing; maybe the way academia, industry, and government interact, and how we develop software will change too. 

**Gastón Milano** is Chief Technology Officer at GeneXus in Uruguay.

**Diego Vallespir** is an adjunct professor at the Universidad de la República in Uruguay.

**Alfredo Viola** is a professor at the Universidad de la República in Uruguay.

Gastón Milano participated actively as chief architect in the development of coronavirus.u.y. Diego Vallespir and Alfredo Viola did not participate in the development of the coronavirus.u.y system.

© 2020 ACM 0001-0782/20/11



# Contextualized Interpretable Machine Learning for Medical Diagnosis

BY WAGNER MEIRA JR., ANTONIO L.P. RIBEIRO, DERICK M. OLIVEIRA, AND ANTONIO H. RIBEIRO

**T**HE EVOLUTION of artificial intelligence and related technologies have the potential to drastically increase the clinical importance of automated diagnosis tools. Putting these tools into use, however, is challenging, since the algorithm outcome will be used to make clinical decisions and wrong predictions can prevent the most appropriate treatment from being provided to the patient. Models should not only provide accurate predictions, but also evidence that supports the outcomes, so they can be audited, and their predictions double-checked. Some models are constructed in such a way they are difficult to interpret, hence the name *black-box models*. While there are methods that generate explanations for generic black-box classifiers,<sup>9</sup> the

solutions are usually not tailored for the needs of physicians and do not take any medical background into consideration. Our claim, in this work, is that explanations must be based on features that are meaningful to physicians. We call those contextual features.

Deep neural networks are relevant examples of black-box models. These models, trained on large real datasets, have demonstrated the ability to provide extremely accurate diagnosis.<sup>1,5</sup> However, these large and complex models of stacked transformations usually do not allow easy interpretation of the results. Despite their potential to transform healthcare and clinical practice,<sup>3,8</sup> there are still significant challenges that must be addressed. For instance, it is commonplace that neural network results are brittle either because it learns to solve the task in



unwanted ways or because even small perturbations may have a huge impact on its outcome.<sup>2</sup>

Cardiovascular diseases are the leading cause of death worldwide<sup>7</sup> and the electrocardiogram (ECG) is a major exam for screening cardiovascular diseases (see Figure 1). Our immediate application scenario is the Telehealth Network of Minas Gerais (TNMG), that serves more than 1,000 remote municipalities in six Brazilian states. More than 2,000 ECGs are examined daily and reported by cardiologists using a Web-based system. Our goal is to empower those physicians through not only accurate, automatically generated disease predictions, but also

explanations that ease their understanding of the model outcome.

Classical methods for automated ECG analysis, such as the University of Glasgow ECG analysis program,<sup>4</sup> employ a two-step approach: First extracting the main features of the ECG signal using traditional signal processing techniques and then using these features as inputs to a classifier. Deep learning presents an alternative to this approach, since the raw signal itself is used as an input to the classifier, which learns from examples to extract the features, as presented in our previous work.<sup>6</sup> In the classical two-step approach, the models are built on top of measures and features

**In order to improve accuracy and transparency in automatic ECG analysis, we propose generating explanations based on contextual features for ECG diagnosis.**



that are known by the physicians, making it easier to verify and to understand the algorithm decisions as well as to identify sources of algorithmic mistakes. Such transparency is lost in “end-to-end” deep learning approaches.

In order to improve accuracy and transparency in automatic ECG analysis, we propose generating explanations based on contextual features for ECG diagnosis (Figure 2). To the best of our knowledge, this is the first work that generates explanations tailored to physicians’ needs for ECG black-box algorithms, including end-to-end classification models. The proposed method (Figure 3) uses a noise-insertion strategy to quantify the impact of the ECG intervals and segments on the automated classification outcome and to generate meaningful features to the user. These

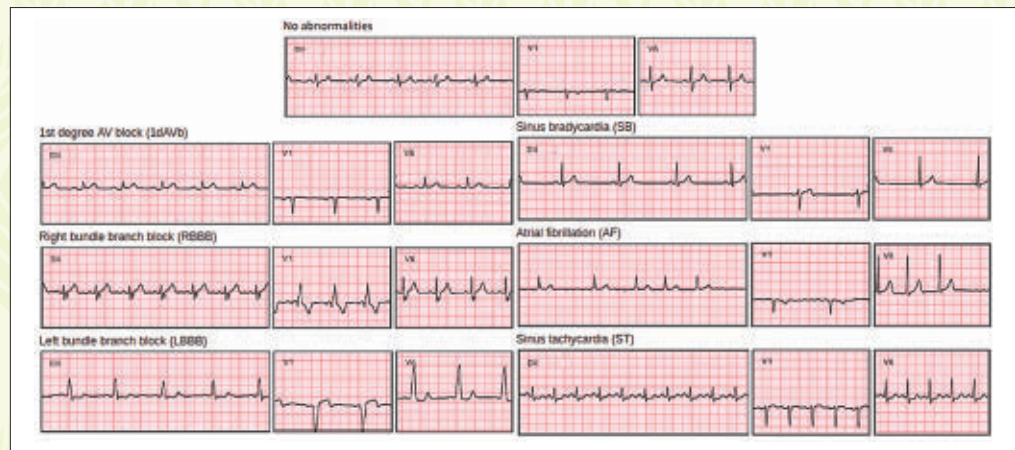


Figure 1. ECG samples for some common diseases.

intervals and segments and their impact on the diagnosis are commonplace to cardiologists, and their usage in explanations enables a better understanding of the outcomes and also the identification of sources of mistakes. We applied our method to generate an explanation to the predictions of the deep learning model presented in Ribeiro et al.<sup>6</sup> using data from TNMG.

Finally, we assessed our approach by analyzing the explanations generated in terms of their interpretability and robustness.

While diagnosing some diseases, cardiologists analyze the ECG (depicted in Figure 4) and apply rules to diagnosis. For instance, the criteria for Left Bundle Branch Block (LBBB) is: QRS duration greater than 120 milliseconds; absence

of Q wave in leads I, V5 and V6; monomorphic R wave in I, V5 and V6; and ST and T wave displacement opposite to the major deflection of the QRS complex. Our explanation consists of both a textual and a visual component in order to better explain to cardiologists in terms and criteria familiar to them. In Figure 5, we show an explanation for six classes of diseases based

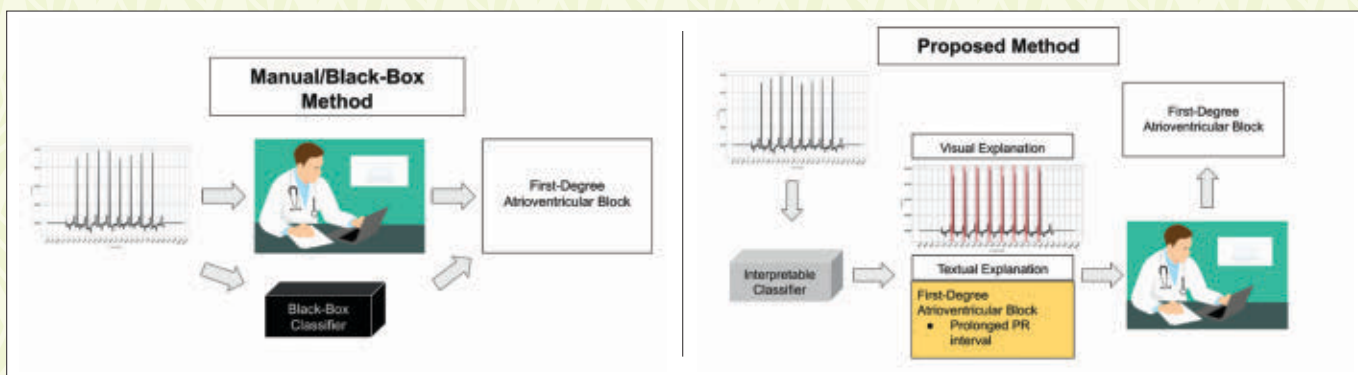


Figure 2. Comparison between methods.

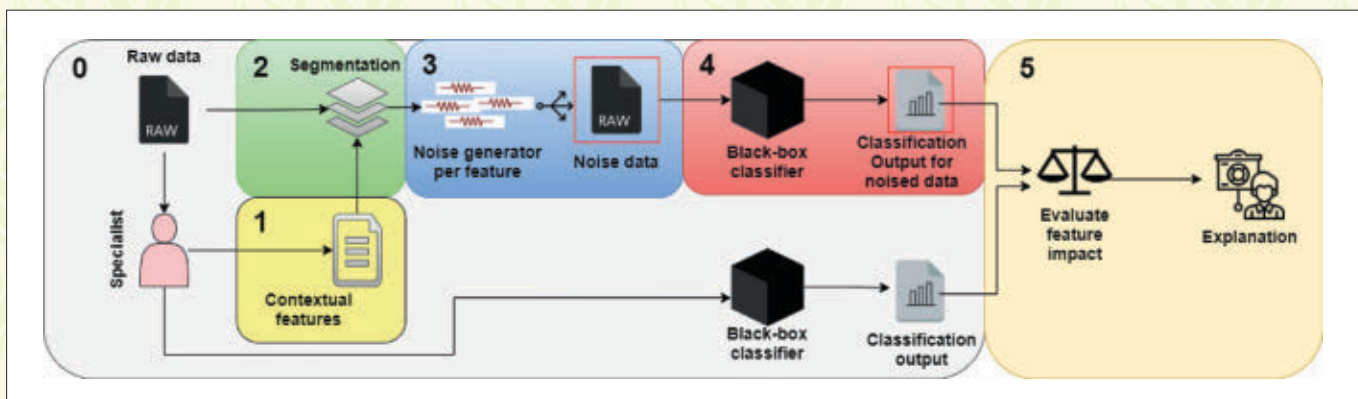


Figure 3. Methodology.

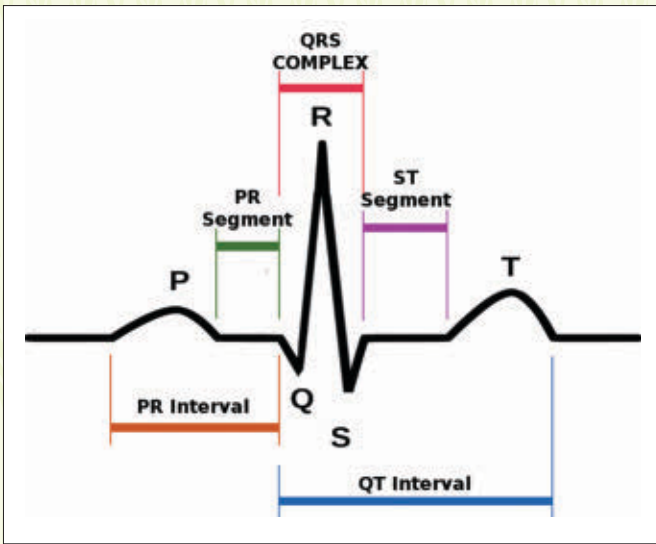



Figure 4. ECG-based diagnosis.

on how much impact the noise has over the different features, quantifying how the different criteria affect the model predictions.

In summary, improving transparency and accountability of deep learning models is an important step toward utilization. Incorporating such models in the TNMG pipeline may improve the quality of its service and have a positive impact in the treatment of many patients. In countries such as Brazil, where the population is spread across large portions of the territo-

ry and access to physicians, in particular specialists, is still an issue, we believe our proposal is an example of research-intensive work that opens new opportunities for the massive and responsible adoption of social impacting initiatives.

**Acknowledgment.** This work is partially supported by the Brazilian agencies CNPq, CAPES and Fapemig, by the projects MASWEB, INCT-Cyber and Atmosphere, and by the Google Research Awards for Latin America program. 

**References**

1. Bejnordi, B.E. et al. Diagnostic assessment of deep Learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 318, 22 (Dec. 2017), 2199; <https://doi.org/10.1001/jama.2017.14585>.
2. Goodfellow, I.J., Shlens, J. and Szegedy, C. Explaining and Harnessing Adversarial Examples, Dec. 2014; arXiv:1412.6572.
3. Hinton, G. Deep learning—A technology with the potential to transform health care. *JAMA* 320, 11 (Sept. 2018), 1101–1102; <https://doi.org/10/gfkh6>.
4. Macfarlane, P.W., Devine, B. and Clark, E. The University of Glasgow (Uni-G) ECG Analysis Program. *Computers in Cardiology* (2005), 451–454; <https://doi.org/10.1109/CIC.2005.1588134>
5. McKinney, S.M. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (Jan. 2020), 89–94; <https://doi.org/10.1038/s41586-019-1799-6>
6. Ribeiro, A.H. et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Commun.* 11 (2020).
7. Ribeiro, A.L.P., Duncan, B.B., Brant, L.C.C., Lotufo, P.A., Mill, J.G. and Barreto, S.M. Cardiovascular health in Brazil: Trends and perspectives. *Circulation* 133, 4 (2016), 422–433.
8. Topol, E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hachette, U.K., 2019.
9. Zaki, M.Z. and Meira Jr., W. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (2<sup>nd</sup> ed.). Cambridge University Press, 2020.

**Wagner Meira Jr.** is a professor in the Department of Computer Science at the Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

**Antonio L.P. Ribeiro** is a professor in the Department of Medical Clinic at the Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

**Derick M. Oliveira** is a Ph.D. student in the Department of Computer Science at the Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

**Antonio H. Ribeiro** is an associate researcher in the Department of Computer Science at the Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

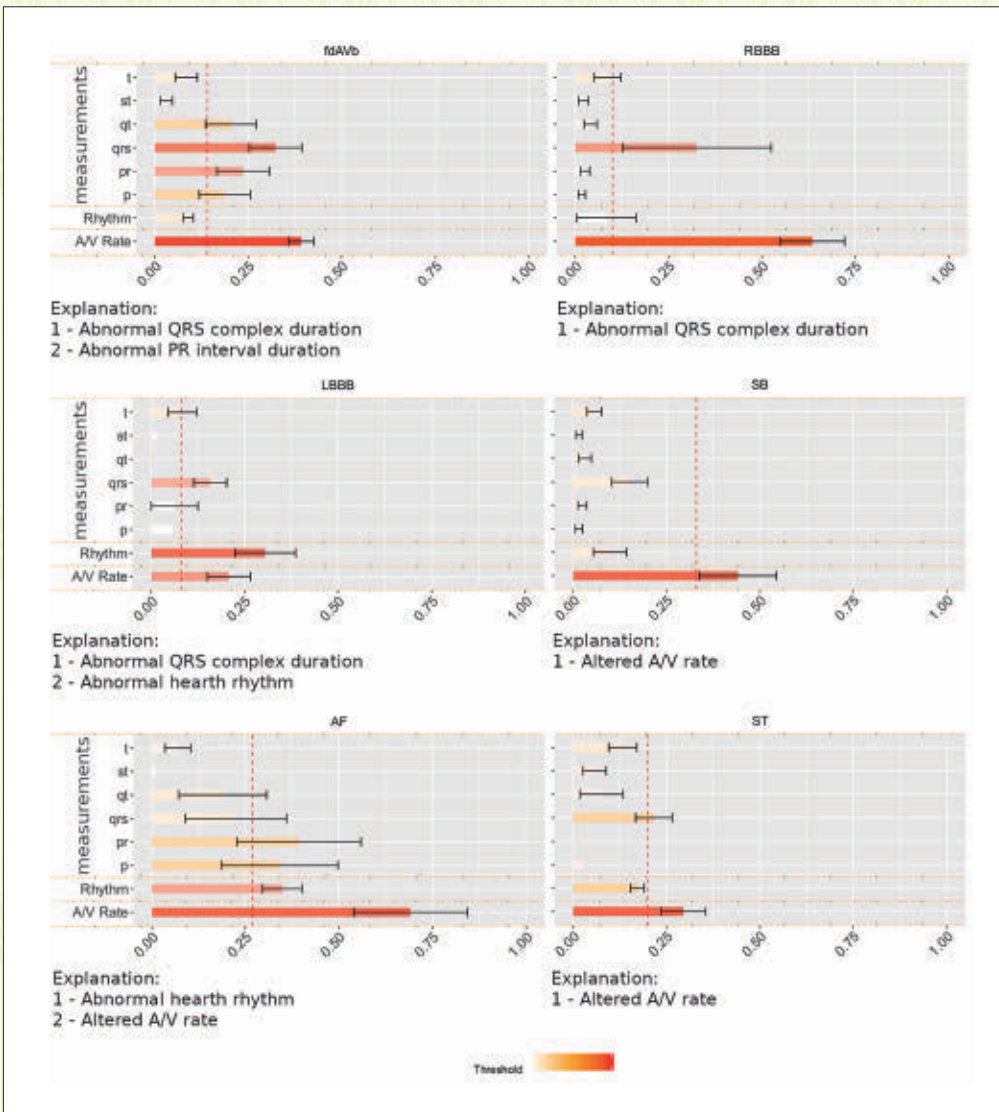


Figure 5. Each explanation has a visual and textual component. The visual component is a horizontal bar graph where each bar represents a feature. The colored bar is the mean value of the impact of the associated feature on the classifier and the error bar at the right end is the standard deviation. An explanation is significant when the mean and the standard deviation are above the threshold vertical dotted line. The textual component is generated automatically.



# Understanding Salsa: How Computing Is Defining Latin Music

BY CARLOS ARCE-LOPERA AND GERARDO M. SARRIA M.

**L**ATIN AMERICA, WITH its rich and varied cultural heritage, is a region widely known by its diverse musical rhythms. Indeed, music and dance constitute an important part of Latin American cultural assets and identity.<sup>2</sup> Some of these rhythms, although famous worldwide, belong to specific regions; for example, samba is from Brazil, tango is from Argentina, merengue is from the Dominican Republic, corrido is from Mexico and vallenato is from Colombia, among many other examples. Most of them were created by the cultural interaction between people from African, Native American, and European cultures that shared their music and instruments. Those heterogeneous cultural characteristics made these music styles appealing to an international audience.

One of the most well-



known Latin musical genres is salsa—a genre considered a fusion between several popular Latin music styles such as son montuno, guaracha, mambo, bolero, and chachacha, to mention a few. However, most of the research on the salsa musical genre is based on ethnomusicology, which

has revealed salsa's complex and diverse social and cultural aspects. Indeed, salsa is an amalgamation of Latin traditions and musical styles representative of the Afro-Hispanic culture in the urban-industrial working class. Salsa songs voice concerns of scarcity, violence, inequality, and desperation, which are musically translated using violent orchestration.<sup>3</sup>

From a computational perspective, there is an interdisciplinary research field called music information retrieval (MIR) that seeks to develop computational data search and retrieval techniques applied to music. MIR combines various engineering techniques such as signal processing, informatics,

machine learning, and computational intelligence with the results of careful experimentation in psychoacoustic and musicology research. MIR focuses on developing a variety of efficient computational algorithms; for example, for genre or instrument recognition, for temporal pattern detection, or even for song structure detection and composition. These state-of-the-art algorithms enable the discovery, organization, and monetization of continuously evolving and growing media collections.

However, most of the research has been focused on popular western music such as pop, rock, and classical music, which usually follow simple and regular

**Computational research on intangible cultural heritage contributes to the legitimization of local music as a valuable part of the identity of a region.**



## Latin American researchers created the Salsa Dataset with more than 20,000 salsa songs, containing metadata and acoustic features, including the analysis of the entire wave for each song.

melodic and temporal patterns. Unfortunately, ethnic and folk non-western music from Asia, Africa, and Latin America has been underrepresented in the MIR community<sup>9</sup> by grouping them into a single category labeled World Music. This particular oversimplification undermines the diversity, richness, and complexity of music in these regions. In particular, the salsa musical genre was not included in rigorous musical research due to two main reasons: the unavailability of salsa musical information on music research datasets and its complex heterogeneity. To address these difficulties, recent MIR research is beginning to include salsa songs in research datasets. In 2008, the Latin Music Database<sup>8</sup> did the first attempt by sharing 311 salsa songs. Also, the Proud Music record label<sup>a</sup> provided 66 entries of salsa songs with features such as tempo, mood, character, instruments, arrangement, and composer.

Nearly a decade later, Latin American researchers created the Salsa Dataset<sup>6,7</sup> with more than 20,000 salsa songs, containing meta-

data and acoustic features, including the analysis of the entire wave for each song. Later, the dataset was modified to contain only 10 seconds of each song in order to improve its shareability. Those 10 seconds corresponded to the most representative part of the song, which in salsa is the chorus. To find the chorus of each song, a chorus extraction algorithm that follows a sequence alignment process was used.<sup>1</sup> This type of algorithm is widely used in bioinformatics to find repeated chains of amino acids. In this sense, by replacing the nomenclature of DNA ACTG (C: cytosine, G: guanine, A: adenine and T: thymine) with the 12 musical notes (C, C#, D, D#, E, F, F#, G, G#, A, A#, B), the sequence alignment algorithm could search similarities between several parts of the song until the chorus was found. This approach achieved a 74% accuracy in chorus detection and revealed that the mambo (the instrumental part of a salsa song) has a similar behavior with respect to the chorus.<sup>1</sup>

The Salsa Dataset was further modified by conducting an analysis to understand the topology of the songs.<sup>6</sup> The analysis, performed using unsupervised machine learning

techniques such as  $k$ -means and self-organizing maps, yielded unsatisfactory results, meaning that the dataset needed an extension that included annotations. The annotations were done manually by an expert in the salsa genre. The expert rated nearly 10,000 10-second audio fragments by associating the audio with its ability to enable dancing the way salsa songs are danced. This rating resulted in a percentage of salsa for each song fragment. Additionally, the expert identified different genres in the audio fragment (from a list of 40 musical genres and subgenres typically related to salsa, such as salsa itself, guaracha, son Cubano, bolero son, bolero tradicional, son montuno, pachanga, merengue and guajira). By combining these multi-labeled annotations with the acoustic features of the song, the Salsa Dataset can provide a complex testing ground for supervised machine learning algorithms.<sup>6</sup>

Besides music classification, another main task in MIR is the automatic composition of music. A formal system based on probabilistic parsers<sup>5</sup> automatically generated salsa music inspired by songs from the Grupo Niche,<sup>4</sup> a famous Colombian salsa band. A Web-based tool to visualize and listen to the possible outcomes of the model was also implemented. This tool allowed the recognition of the patterns used by the band in their songs with the bass and piano scores of the generated song.

Computational research on intangible cultural heritage contributes to the legitimation of local music as a valuable part

of the identity of a region. Furthermore, the proliferation of audio streaming platforms has encouraged studies concerning the automatic classification of musical genres and the successful development of recommender systems for a global audience. The inclusion of the salsa musical genre on the MIR research community represents an opportunity for Latin American researchers to showcase their culture using computational techniques and formal models. 

### References

1. Arévalo, C., Sarria M, G.M., Mora, M., and Arce-Lopera, C. Towards an efficient Algorithm to get the chorus of a salsa song. In *Proceedings of the 2015 IEEE Intern. Symp. Multimedia*, 258–261; <https://doi.org/10.1109/ISM.2015.42>
2. Brill, M. *Music of Latin America and the Caribbean*. Routledge, 2016.
3. Duany, J. Popular music in Puerto Rico: Toward an anthropology of "Salsa." *Latin American Music Review / Revista de Música Latinoamericana* 5, 2 (Oct. 1984), 186–216; DOI: <https://doi.org/10.2307/780072>
4. Ossa, L. Conciencia social y la herencia africana en la salsa de Joe Arroyo y Grupo Niche. *Afro-Hispanic Review* 23, 2 (2004), 62–69.
5. Rodríguez, B., de Piñérez, R.G. and Sarria M, G.M. 2017. Using probabilistic parsers to support salsa music composition. *Mathematics and Computation in Music (LNCS)*, Springer International Publishing, 361–372; [https://doi.org/10.1007/978-3-319-71827-9\\_28](https://doi.org/10.1007/978-3-319-71827-9_28)
6. Sarria M, G.M., Diaz, J. and Arce-Lopera, C. Analyzing and extending the salsa music dataset. In *Proceedings of the 2019 XXII Symp. Image, Signal Processing, and Artificial Vision*, 1–5; <https://doi.org/10.1109/STISIVA.2019.8730229>
7. Sarria M, G.M., Mora, M. and Arce-Lopera, C. 2016. Salsa dataset: primera base de conocimiento de música salsa. *Ricercare* 5 (Dec. 2016), 63–72; <https://doi.org/10.17230/ricercare.2016.5.5>
8. Silla, C.N., Koerich, A.L. and Kaestner, C.A.A. *The Latin Music Database* (2008), 451–456.
9. Sturm, B.L. A survey of evaluation in music genre recognition. *Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation (LNCS)*. Springer International Publishing, 2014, 29–66; [https://doi.org/10.1007/978-3-319-12093-5\\_2](https://doi.org/10.1007/978-3-319-12093-5_2)

Carlos Arce-Lopera is an associate professor at Universidad Icesi in Cali, Colombia.

Gerardo M. Sarria M. is an associate professor and the Head of the Computer Science Program at Pontificia Universidad Javeriana Cali, Colombia.

a Royalty free music tracks, genre Salsa 1/3; <https://www.proudmusiclibrary.com/en/genre/salsa>

# Minding the AI Gap in LATAM

BY BARBARA POBLETE AND JORGE PÉREZ

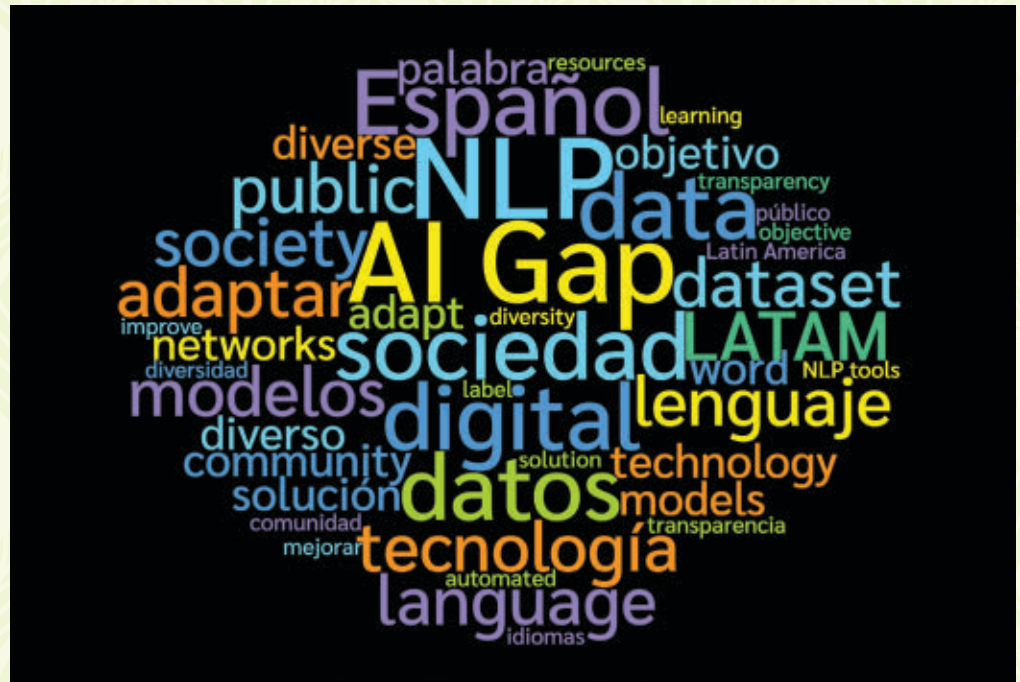
**S**Oieties and industries are rapidly changing due to the adoption of artificial intelligence (AI) and will face deep transformations in upcoming years. In this scenario, it becomes critical for under-represented communities in technology, in particular developing countries like Latin America, to foster initiatives that are committed to developing tools for the local adoption of AI. Latin America, as well as many non-English speaking regions, face several problems for the adoption of AI technology, including the lack of diverse and representative resources for automated learning tasks. A highly problematic area in this regard is natural language processing (NLP), which is strongly dependent on labeled datasets for learning. However, most state-of-the-art NLP resources are allocated to English. Therefore, creating efficient NLP tools for diverse languages requires

an important investment of time and financial resources. To deal with such issues, our group has worked toward creating *language-agnostic approaches* as well as adapting and improving existing NLP techniques to local problems. In addition, we have focused on producing new state-of-the-art NLP publicly available data and models in Spanish. Next, we briefly

present some of them.

**Twicalli, a social seismograph, and other crisis management tools.** Timely detection and accurate description of natural disasters and other crisis situations are crucial for emergency management. This is challenging and important for our region, since one must rely on human observers appointed to specific geographical areas or on advanced infrastructure. In the case of earthquakes, geographically dense sensor networks are expensive. A viable inexpensive alternative to this problem is to detect events through people's reactions in online social networks, particularly on Twitter.<sup>a</sup>

Nevertheless, the massive number of messages in the data stream, along with the noise they contain, create a number of difficulties for worldwide detection. The common solution used to date has been to learn from labeled data to identify user messages related to a real-time earthquake.<sup>2,12</sup> This approach does not scale well globally across countries and languages. Consequently, our group proposed a simple, yet efficient solution that, generally speaking, identifies the unusual increase in the frequency of multilingual textual features related to earthquakes within the Twitter stream.<sup>10</sup> This method only requires a one-off semi-supervised initialization and can be scaled to track multiple features and thus, multiple



**Our group has worked toward creating language-agnostic approaches as well as adapting and improving existing NLP techniques to local problems.**

<sup>a</sup> Microblogging and social networking service <http://twitter.com>



crisis situations. Experimental results validate our approach as a competitive open source alternative to leading solutions, with the advantage of working independently of language and providing worldwide scalability. An instance of this framework is currently available as *Twicalli* (<http://twicalli.cl>) that provides visual “social seismograph” for the Chilean territory (as shown in Figure 1). *Twicalli* is used by the National Seismology Center in Chile, among other emergency response agencies nationwide. Furthermore, we are in the process of incorporating a novel machine learning-based model for automatically estimating the Modified Mercalli Intensity Scale<sup>b</sup> of an earthquake,<sup>9</sup> and detecting in real time other types of crisis situations.<sup>13</sup>

**Political data and the Open Constitution Project.** In 2016, Chile went through a collective open process to establish the guidelines of what a new political constitution should consider. From a



**Figure 1.** Twicalli interface showing a large earthquake in Chile on December 25, 2016 and its aftershocks. (Top right) Shows the frequency of earthquake-related messages per minute. (Left) Shows a heat map of geographical message density distribution. (Bottom center) Displays user messages. (Bottom right) Shows fine-grained geographical message navigation. Image courtesy of Jazmine Maldonado.

b <https://www.usgs.gov/natural-hazards/earthquake-hazards/science/modified-mercalli-intensity-scale>

technological point of view, an interesting aspect of the process was the use of digital platforms to collect

**With the help of an interdisciplinary group of linguists and experts in argumentation, we first determined the most important tasks and then developed machine learning methods to solve them.**

the output of the deliberative instances that included discussions produced by 8,000+ small assemblies across the country. This resulted in a dataset of 200,000+ political arguments, which was openly published in a raw and anonymous form. This dataset was manually systematized through months of work. Hence, with the goal of making this process less time consuming, more objective, and scalable, we worked toward creating ways to automate at least parts of the systematization. With the help of an interdisciplinary group of linguists and experts

in argumentation, we first determined the most important tasks and then developed machine learning methods to solve them. For example, we addressed the task of classifying raw text arguments into the corresponding “constitutional concept.” For instance, the raw text “*the state should provide free education for all*” should be assigned to the concept “*right to education*.” This was challenging, given we had more than 100 different constitutional concepts. Our best method achieved a top-5 accuracy of more than 90%.<sup>6</sup> We created a visualization for exploring the dataset



(<http://constitucionabierta.cl/>), which was widely used by the public and press, providing a much needed transparency to such an important process (see Figure 2).


Two problems naturally arise in the context of our work analyzing social and political discussions; that of incivility, or hate speech,<sup>1,7</sup> and the now ubiquitous problem of bias against minorities.<sup>3,8</sup> Modern automatic tools for processing human-generated text should consider these issues as essential. However, these tasks are extremely challenging even for the English language. Specifically, our work has addressed how the lack of diversity in training data for hate-speech detection models induces an over-estimation of their performance in state-of-the-art approaches, and how this affects transferring this knowledge to other domains,<sup>1</sup> such as Spanish.

### Spanish NLP resources.

An issue we have constantly faced, as have other

researchers working with Spanish text data, is the lack of high-quality resources for developing, training, and testing models. Some LATAM and Spanish groups have been tackling this problem by systematically producing freely available NLP resources for the whole research community.<sup>4,5,11,14,15</sup> One set of resources of particular importance are Word Embeddings trained from big corpora.<sup>5,14,15</sup> Our group has also developed a Spanish pretrained model based on Neural Self-Attention<sup>4</sup> that has improved the state of the art in many NLP tasks in Spanish. We hope more groups in Latin America can join efforts to produce resources to improve NLP research and applications.

### Acknowledgments.

BP and JP were partially funded by Fondecyt grants 1191604 and 1200967, respectively. 

### References

1. Arango, A., Pérez, J. and Poblete, B. Hate speech detection is not as easy as you may think: A closer look at model validation. *SIGIR* 2019: 45–54.

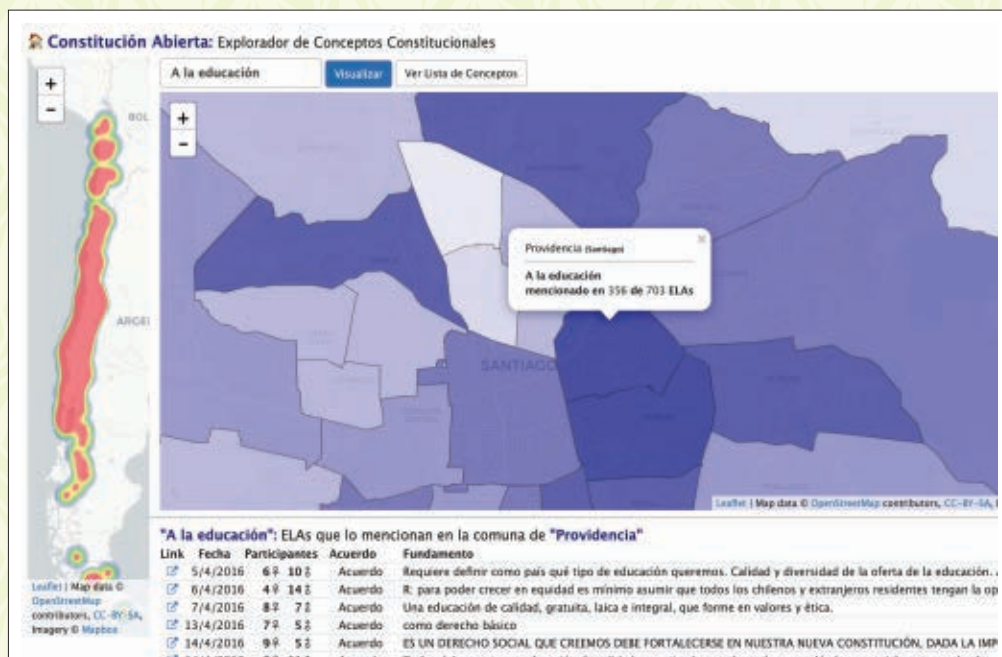
## An issue we have constantly faced, as have other researchers working with Spanish text data, is the lack of high-quality resources for developing, training, and testing models.

2. Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C. and Tesconi, M. (2014, August). "EARS (earthquake alert and report system) a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, Aug. 2014, 1749–1758.
3. Badilla, P., Bravo-Marquez, F. and Pérez, J. WEF: The word embeddings fairness evaluation framework. To appear in *Proceedings of the 29th Intern. Joint Conf. Artificial Intelligence*, 2020.
4. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H. and Perez, J. Spanish pre-trained BERT model and evaluation data. In *Practical Machine Learning for Developing Countries at ICLR 2020*; <https://github.com/ccuchile/beto>
5. Etcheverry, M. and Wonever, D. Spanish word vectors from Wikipedia. In *Proceedings of 2016 Intern. Conf. Language Resources and Evaluation*.
6. Fierro, C., Fuentes, C., Pérez, J. and Quezada, M. 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, Sept. 2017, 1–10.
7. Fortuna, P. and Nunes, S. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.* 51, 4 (2018), 85:1–85:30.
8. Garg, N., Schiebinger, L., Jurafsky, D. and Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. In *Proceedings of the National Academy of Sciences*, 115, 16 (2018), E3635–E3644.
9. Mendoza, M., Poblete, B. and Valderrama, I. Nowcasting earthquake damages with Twitter. *EPJ Data Sci.* 8, 3 (2019). <https://doi.org/10.1140/epjds/s13688-019-0181-0>
10. Poblete, B., Guzmán, J., Maldonado, J. and Tobar, F. Robust detection of extreme events using Twitter: Worldwide earthquake monitoring. *IEEE Trans. Multimedia* 20, 10, (Oct. 2018), 2551–2561; doi: 10.1109/TMM.2018.2855107.
11. Recursos, Grupo de Procesamiento de Lenguaje Natural, Universidad de La República; <https://www.fing.edu.uy/inco/grupos/pln/recursos.html>
12. Sakaki, T., Okazaki, M. and Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th Intern. Conf. World Wide Web*, Apr. 2010, 851–860.
13. Sarmiento, H., Poblete, B. and Campos, J. Domain-independent detection of emergency situations based on social activity related to geolocations. In *Proceedings of the 10th ACM Conf. Web Science*, 2018; <https://doi.org/10.1145/3201064.3201077>
14. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M. and Armengol-Estapé, J. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop at ACL 2019*.
15. Spanish Word Embeddings; <https://github.com/ccuchile/spanish-word-embeddings>

**Barbara Poblete** is an associate professor in the Department of Computer Science at Universidad de Chile and an associate researcher at the Millennium Institute for Foundational Research on Data, Santiago, Chile.

**Jorge Pérez** is an associate professor in the Department of Computer Science at Universidad de Chile and an associate researcher at the Millennium Institute for Foundational Research on Data, Santiago, Chile.

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



**Figure 2.** Interface of the website <http://constitucionabierta.cl> showing the distribution of the constitutional concept “right to education” in the 2016 Chilean Constitutional discussions dataset.



# Three Success Stories About Compact Data Structures

BY DIEGO ARROYUELO, JOSÉ FUENTES-SEPÚLVEDA, AND DIEGO SECO

**T**ECHNOLOGY EVOLUTION IS no longer keeping pace with the growth of data. We are facing problems storing and processing the huge amounts of data produced every day. People rely on data-intensive applications and new paradigms (for example, edge computing) to try to keep computation closer to where data is produced and needed. Thus, the need to store and query data in devices where capacity is surpassed by data volume is routine today, ranging from astronomy data to be processed by supercomputers, to personal data to be processed by wearable sensors. The scale is different, yet the underlying problem is the same.

Keeping data structures in fast memory has been the major workhorse of compact data structures (CDS), with enormous practical benefits, such as speeding up data processing and querying.<sup>10</sup> This has situated CDS at the forefront of research in data

structures over the last 20 years. Remarkable contributions have been made, many of them by Latin American researchers, forming an active and prolific community spread mainly over three Chilean universities, and branching from there to other universities in Chile and Latin America as well as to Europe, the U.S., Canada, East Asia, and Australia. The senior researcher at Universidad de Chile is Gonzalo Navarro; the groups at Universidad Técnica Federico Santa María and Universidad de Concepción are integrated by Diego Arroyuelo, José Fuentes-Sepúlveda, and Diego Seco. Here, we describe three success stories with strong roots in the region.

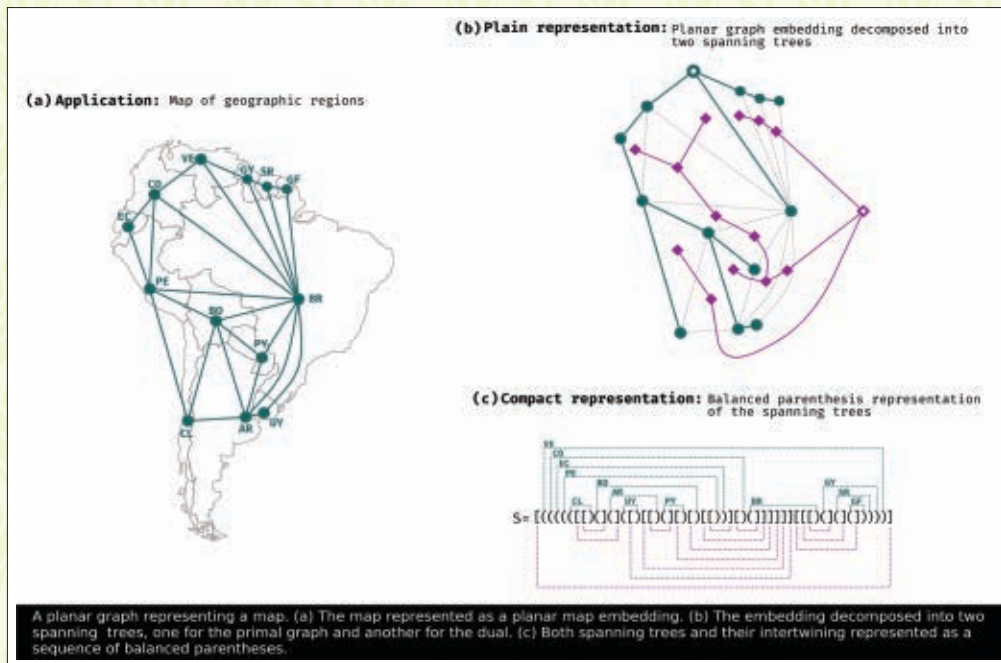
Trees (particularly ordinal, cardinal, and binary trees) are a paradigmatic example of CDS success. After Jacobson's seminal work,<sup>9</sup> several ordinal tree CDS were presented, most achieving the  $2n$ -bit lower bound for an  $n$ -node tree: only 2-bits per node, up to 32x smaller than a pointer-based representation!

Within this bound, researchers supported, progressively, more tree operations. The most outstanding result in this line supports a striking set of over 25 operations in constant time for static ordinal trees (sublogarithmic for dynamic trees).<sup>11</sup> The key was the ability to translate all tree operations into operations on a sequence of  $2n$  balanced parentheses (as former approaches), which are in turn supported by a single data structure of sublinear ( $o(n)$ -bit) additional space: the so-called range min-max tree.<sup>11</sup> This represented remarkable progress, as former approaches typically needed a separate data structure per operation, which is impractical. Indeed, this approach is practical: about 2.1-bits per tree node on average,<sup>1</sup> with operation times within a few microseconds. For dynamic  $k$ -ary cardinal trees, the best result uses space close to the optimal, supporting a set of 13 operations in constant time for  $k = O(\log n)$ <sup>2</sup> (sublogarithmic time for general  $k$ ). Besides its independent theoretical interest, this motivated new results in related areas, for example, the construction of compressed text indexes.<sup>3</sup> For dynamic binary trees, one can use optimal space and support tree operations in constant time (including insertions/deletions), while associating values to tree nodes space-efficiently.<sup>2</sup> Both features were not achieved jointly before. On a related line, the prog-

ress on balanced parenthesis sequences<sup>11</sup> led to compact representations of planar graphs of  $e$  edges using only  $4e + o(e)$  bits,<sup>6</sup> with direct applications in geographic information systems (GIS).<sup>7</sup>

The second success story is that of  $k^2$ -trees,<sup>4</sup> a compact representation of well-known quadtrees. Although originally designed to support fast navigation on Web graphs, their versatility allows one to support operations like variants of range queries on grid points, and use them in domains such as binary relations, RDF databases, raster and trajectory data on GIS, and OLAP cubes in data warehouses. Even though  $k^2$ -trees do not provide good theoretical guarantees, they usually perform well in practice. In a nutshell, they exploit large empty areas that arise in the adjacency matrix of usual graphs (or any binary matrix, in general), requiring less space when the binary matrix is clustered. For Web graphs, they require 1–3-bits per graph edge (that is, per non-empty entry in the matrix), while supporting usual graph operations efficiently. A remarkable contribution was the support of direct and reverse neighbors within such space, whereas previous approaches needed to double the space. Conceptually, a  $k^2$ -tree is obtained by recursively partitioning the matrix into  $k^2$  submatrices, which for  $k=2$  resembles the quadtree partitioning.

**The need to store and query data in devices where capacity is surpassed by data volume is routine today.**



### South American map graph.


The tree obtained from this hierarchical subdivision is represented level-wise using bitmaps, avoiding the use of pointers. Operations on the tree are simulated using fast operations on the bitmaps. This also facilitates the addition of data to the tree nodes, which is used to augment the basic structure with, for example, integer values of a raster or OLAP cube. k2-trees have been used recently to support worst-case optimal joins in graph databases.<sup>12</sup>

The third success story regards the synergy between CDS and bioinformatics,<sup>5</sup> evidenced by an international collaboration project,<sup>8</sup> where Latin American researchers played a central role. In this context, the research on indexing highly repetitive text collections has provided outstanding results. These collections can be found naturally in genomic databases (storing genome sequences of individuals of the same

species), versioned text collections (for example, Wikipedia), and software repositories (for example, GitHub). Classical compressed text indexes, such as FM-indexes, provide search functionalities using space close to the statistical entropy of the text collection, being insensitive to repetitiveness. Recently, a novel index, called  $r$ -index,<sup>8</sup> was proposed to exploit text repetitiveness. It uses space close to  $r$ , the number of runs (maximal substring consisting of a single character) of the Burrows-Wheeler Transform<sup>10</sup> of the text. A highly repetitive text yields a small  $r$ . Within  $O(r)$  space, the  $r$ -index supports a complete set of search functionalities in  $O(\log(n/r))$  time,  $n$  being the text length. Compared to the state-of-the-art short-read aligner Bowtie,<sup>9</sup> the  $r$ -index typically uses less than 0.1 bits per base, whereas Bowtie uses 4-bits per base, both having comparable running times.

The  $r$ -index is expected to be a breakthrough in the development of a new generation of space-efficient bioinformatic tools.

Today, space-efficient algorithms and data structures have become a must. Initially unsuspected results have been accomplished over the years. The introduction of the Internet of Things, scientific initiatives (for example, particle colliders and astronomical observatories), seismic/volcano monitoring systems, edge computing (which involves semi-autonomous devices with limited memory), and smartphone applications, among others, impose challenges to CDS. For instance, a new observatory in northern Chile is expected to generate 20 terabytes of data every night. This challenges the ability to build CDS for huge amounts of data and handle data in streaming mode (so it can be queried as it is received). Similarly, supporting multi-terabyte catalogs of satellite imagery and geospatial datasets, like the one by Google Earth Engine, is a challenge

for raster representations based on CDS. Finally, since CDS aim at reducing memory transfers, they might help to design energy-aware algorithms. The maturity of the field will be key to face the data flood. 

### References

1. Arroyuelo, D., Cánovas, R., Navarro, G. and Sadakane, K. Succinct trees in practice. In *Proceedings of the Workshop on Algorithm Engineering and Experiments*. SIAM, 2010, 84–97.
2. Arroyuelo, D., Davoodi, P. and Rao Satti, S. Succinct dynamic cardinal trees. *Algorithmica* 74, 2 (2016), 742–777.
3. Arroyuelo, D. and Navarro, G. Space-efficient construction of Lempel-Ziv compressed text indexes. *Information and Computation* 209, 7 (2011), 1070–1102.
4. Brisaboa, N.R., Ladra, S. and Navarro, G. Compact representation of web graphs with extended functionality. *Information Systems* 39 (2014), 152–174.
5. Cunial, F., Mäkinen, V., Belazzougui, D. and Tomescu, A.I. *Genome-Scale Algorithm Design—Biological Sequence Analysis in the Era of High-throughput Sequencing*. Cambridge University Press, 2015.
6. Ferrer, L., Fuentes-Sepúlveda, J., Gagie, T., He, M. and Navarro, G. Fast and compact planar embeddings. *Computational Geometry* 89 (2020), 101630.
7. Fuentes-Sepúlveda, J., Navarro, G. and Seco, D. Implementing the topological model succinctly. *String Processing and Information Retrieval*. Springer Intern. Publishing, 2019, 499–512.
8. Gagie, T., Navarro, G. and Prezza, N. Fully functional suffix trees and optimal text searching in BWT-runs bounded space. *J. ACM* 67, 1 (Jan. 2020).
9. Jacobson, G. Space-efficient static trees and graphs. In *Annual Symp. Foundations of Computer Science*. IEEE Computer Society, 1989, 549–554.
10. Navarro, G. *Compact Data Structures—A Practical Approach*. Cambridge University Press, 2016.
11. Navarro, G. and Sadakane, K. Fully functional static and dynamic succinct trees. *ACM Transactions on Algorithms* 10, 3 (2014), 16:1–16:39.
12. Navarro, G., Reutter, J.L. and Rojas-Ledesma, J. Optimal joins using compact data structures. In *Proceedings of the Intern. Conf. Database Theory*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020, 21:1–21:21.

**Diego Arroyuelo** is an assistant professor at the Universidad Técnica Federico Santa María, and the Millennium Institute for Foundational Research on Data, Santiago, Chile.

**José Fuentes-Sepúlveda** is an assistant professor at the Universidad de Concepción, Chile.

**Diego Seco** is an associate professor at the Universidad de Concepción, and the Millennium Institute for Foundational Research on Data, Santiago, Chile.

a BIRDS Project: Bioinformatics and Information Retrieval Data Structures Analysis and Design; <http://www.birdsproject.eu>.

b Bowtie: An ultrafast memory-efficient short read aligner; <http://bowtie-bio.sourceforge.net/>.



BY ISIDORO GITLER, ANTÔNIO TADEU A. GOMES,  
AND SERGIO NESMACHNOW

# The Latin American Supercomputing Ecosystem for Science

LARGE, EXPENSIVE, COMPUTING-INTENSIVE research initiatives have historically promoted high-performance computing (HPC) in the wealthiest countries, most notably in the U.S., Europe, Japan, and China. The exponential impact of the Internet and of artificial intelligence (AI) has pushed HPC to a new level, affecting economies and societies worldwide. In Latin America, this was no different. Nevertheless, the use of HPC in science affected the countries in the region in a heterogeneous way. Since the first edition in 1993 of the TOP500 list of most powerful supercomputing systems in the world, only Mexico and Brazil (with 18 appearances each) made the list with research-oriented supercomputers. As of June 2020, Brazil was the only representative of Latin America on the list.

HPC represents a strategic resource for Latin American researchers to respond to the economical and societal challenges in the region and to cross-fertilize with researchers in the rest of the world. Nevertheless, the Latin American countries still lag behind other countries in terms of size and regularity of investments in HPC. The table here compares the HPC capacity of the BRICS countries, which together represent almost half of the world population. As a reference, in 2018, South Africa's GDP was 29.1% lower than Argentina's and only 11.2% higher than Colombia's, the two countries in Latin America with largest GDPs after Brazil and Mexico. In spite of the overall picture described here, the landscape of the Latin American HPC ecosystem for science is promising, with many initiatives and outstanding concrete results.

## HPC in Latin America: The Cases of Brazil, Mexico, and Uruguay

Comparing the situation of HPC in three different countries in Latin America helps understanding the region's distinctions, not only in terms of overall capacity, but also in terms of adopted policies for creating and operating this kind of scientific instrumentation systems. The presented examples are representative of other important initiatives in the region, for example, NLHPC in Chile, Tupac in Argentina, SC3UIS in Colombia, and CeNAT in Costa Rica.

In Brazil, the National Laboratory for Scientific Computing (LNCC) is the major player for HPC services to the scientific community. LNCC is a public, interdisciplinary research center with a mission-oriented approach to computational and mathematical modeling and simulation of complex problems. LNCC coordinates a network of 10 HPC centers (SINAPAD) funded by the Brazilian Ministry of Science, Technology and Innovations (MCTI) (see Figure 1). The SINAPAD centers offer resources, training, and scientific portals<sup>9,11</sup> to the Brazilian community.

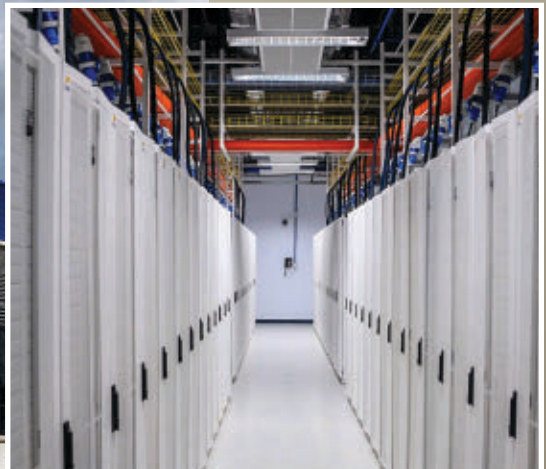




**The Santos Dumont supercomputer at LNCC, Brazil ([www.lncc.br](http://www.lncc.br)).**



**The ABACUS Laboratory at CINVESTAV, Mexico ([www.abacus.cinvestav.mx](http://www.abacus.cinvestav.mx)).**



**Cluster-UY at Massera datacenter, Uruguay ([cluster.uy](http://cluster.uy)).**



**Comparing the situation of HPC in three different countries in Latin America helps understanding the region's distinctions, not only in terms of overall capacity, but also in terms of adopted policies for creating and operating this kind of scientific instrumentation systems.**

LNCC hosts Santos Dumont, which until June 2020 was the largest super-computer dedicated to research in Latin America. With a performance of 4.2 petaflops and a storage capacity of 2.3 petabytes, Santos Dumont plays a central role in promoting high-quality research initiatives in Brazil.

The Santos Dumont supercomputer has housed more than 160 peer-reviewed projects from institutions spread throughout Brazil, which together have consumed more than 500 million CPU hours, produced more than 300 scientific papers in scholarly articles in journals, and supported more than 60 Ph.D. graduations since its inauguration in August 2016. An example of a strategic project on Santos Dumont is the computational modeling of key aspects of the operation of Sirius, the new fourth generation Brazilian synchrotron light source (see <https://www.lnls.cnpm.br/sirius-en/>). Besides, Santos Dumont has fostered important international collaborations, such as in the rational design of Zika vaccine candidates.<sup>7</sup> The LNCC research staff also uses Santos Dumont's capacity to develop high-impact projects, including the development of efficient computational models that quantify the functional severity of coronary artery stenosis,<sup>3</sup> and the genome sequencing, at the beginning of the COVID-19 pandemic, of 19 viruses from different regions of Brazil, demonstrating the state of community transmission.

In Mexico, ABACUS, the Laboratory for Applied Mathematics and HPC at the Center for Research and Advanced Studies (CINVESTAV) exemplifies the diverse HPC initiatives grouped in the Mexican Network in HPC (REDMEX-SU), whose main members are shown in Figure 2.

CINVESTAV is a public institution

ranked within the top Mexican National Research Centers and Postgraduate Education Institutions and through ABACUS and LANCAD a prime provider of HPC resources to the scientific and technological communities in Mexico. CINVESTAV has an outstanding record regarding initiatives to foster interaction between academia, government, industry and society, in conjunction, with a very successful track of worldwide collaboration. ABACUS houses one of the principal Latin American research supercomputers, placed 255th in the TOP500 list of July 2015, with an updated total performance of ~0.5 petaflops and a storage capacity of 1 petabyte.

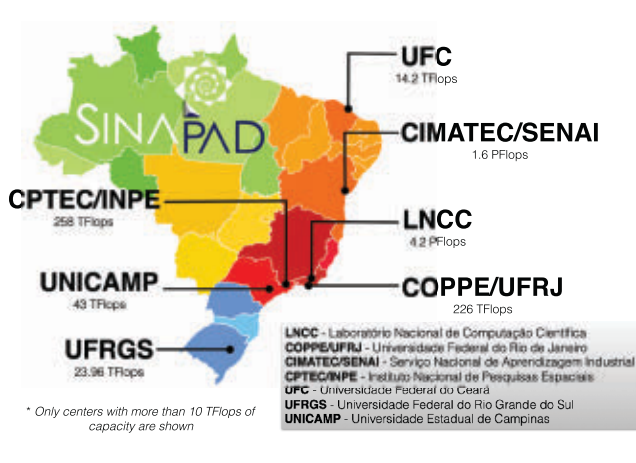
Since 2016, ABACUS has supported scientific efforts to solve complex problems through collaborative work between different research communities spread throughout Mexico. ABACUS has assisted more than 140 research projects and over 250 academic articles. Examples of projects are: numerical simulation of vascular malformations in the brain; studies of racemization of molecular helices; numerical simulation of environmental hazards;<sup>6</sup> sandpile simulations and applications;<sup>5</sup> covering arrays and software testing; cryptographic algorithms; simulation of subatomic processes; simulation of astrophysical phenomena;<sup>4</sup> and the ENERXICO Project: Supercomputing and Energy for Mexico (see <https://enerxico-project.eu>). Some projects have direct impact in daily life and industry, for example, the numerical simulation of volcanic ash dispersion near important airports of the country and early warning systems for coastal floods.

In Uruguay, National Supercomputing Center (Cluster-UY) is an initiative for operating a collaborative scientific HPC infrastructure to foster research

**GDP and investment in supercomputing in the BRICS countries (sources: World Bank, TOP500 List).**

	Nominal GDP 2018 (millions of Dollars)	Entries in Top500 List			
		Nov. '16	Nov. '17	Nov. '18	Nov. '19
<b>Brazil</b>	1,885	3	0	1	3
<b>Russia</b>	1,658	3	3	3	3
<b>India</b>	2,719	5	4	4	2
<b>China</b>	13,608	171	202	227	228
<b>South Africa</b>	368	1	1	2	0

**Figure 1. The main centers of SINAPAD, Brazil**  
([www.lncc.br/sinapad](http://www.lncc.br/sinapad)).



**Figure 2. Main members of the Mexican HPC Network—REDMEXSU**  
([www.redmexsu.mx](http://www.redmexsu.mx)).



and innovation projects with high computing demands. The platform and services provided by Cluster-UY are available to all research and development efforts by scientific institutions, academia, and public/private companies. Important institutions support the initiative, as described in Figure 3.

Cluster-UY applies a collaborative self-managed and self-financed model based on work and support agreements, signed with institutions, organizations, and companies, to guarantee sustainability. Partnerships allow consolidating the collaborative model and are very valuable to open the Center to a wide variety of users, with special emphasis on promoting inclusive development in areas with social impact (health, education, research for development, and so forth), under the “University of Development” model.<sup>8</sup> The approach encourages and consolidates an open data model, closely linked to the ideal of collaborative systems. Furthermore, an egalitarian model is applied for access to the provided services. The same benefits are offered to all users. In turn, all users have the same responsibilities regarding the correct utilization, maintenance, and updating of the platform and services.

Cluster-UY supported more than 50 research projects that used more than 11 million hours, produced more than 250 articles, and supported 100 post-graduate theses since 2018. Relevant projects have been developed using the computing capabilities of Cluster-UY, including: the development of

forecasting tools for renewable energy management in Uruguay (Energy);<sup>10</sup> socioeconomic analysis of short and long-term prices and the impact on welfare of low-income citizens (Economy/Social Sciences); free and publicly available database of biomolecular simulations of SARS-CoV2 proteins (Bioinformatics, Institut Pasteur Montevideo & Cluster-UY);<sup>8</sup> and analysis of dumping of La Teja oil refinery in Montevideo Bay (Environment/Sea Hydraulics/Water quality), among others. These projects have been significant contributions to the aforementioned research lines in Uruguay.

The aforementioned descriptions demonstrate different approaches have been applied for developing and operating HPC facilities in Brazil, Mexico, and Uruguay.

In Brazil, its territory size and federative political model led to different initiatives for fostering HPC in the country. Among them, the most prominent was SINAPAD. Initially devised as a network of HPC centers with similar capacity, the SINAPAD gradually changed to a tiered model with the Santos Dumont supercomputer being its Tier-0.

Mexico has fostered different HPC public infrastructures according to specific regional research and technological needs. As a consequence, at least 10 HPC centers are housed in state universities, federal research centers, and National Council for Science and Technology (CONACYT) centers, among them four with computational capacities similar to ABACUS. These

centers are all members of the Mexican Network in HPC, which offers coordinated HPC resources and training to diverse communities in the country.

Finally, in Uruguay a collaborative model has been applied. The benefits of this model are twofold: on the one hand, it allows implementing an egalitarian model for access to the provided services and incorporating users (scientists, partners, companies) as real owners of the Center; on the other hand, it provides a way to get the much-needed funds for operation, maintenance, and growth in small countries, where funds for research are scarce. This model can be replicated in other small countries in Latin America.

### Networking in Support of Supercomputing Activities

Networking is also a crucial component of collaborative efforts to build an ecosystem for science in Latin America, in particular for sharing HPC infrastructures among its countries. RedCLARA is a Latin American organization created and led by National Research and Education Networks (NRENs), with the main goal of strengthening the development of science, education, culture and innovation in the region through the innovative use of advanced networks (see Figure 4). Created in 2003, RedCLARA led, operated, maintained and developed several infrastructure projects to support research and education (ALICE, ALICE 2, the Bella Program), which allowed the creation of an advanced high-bandwidth net-



Figure 3. Institutions supporting Cluster-UY, Uruguay.

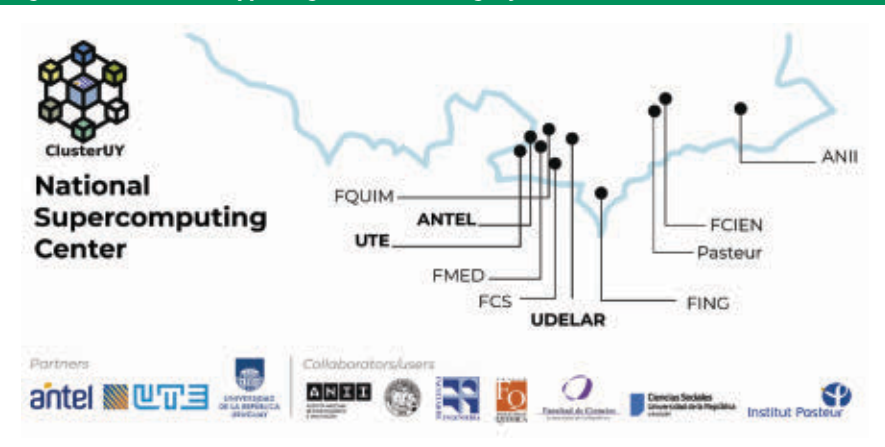


Figure 4. RedCLARA connections (www.redclara.net).



work, interconnecting Latin American NRENs and with NRENs worldwide. RedCLARA also supported other projects (ComGridLatAm, RISC, RICAP) in order to create a continental ecosystem in accordance with the Advanced Computing System for Latin America and Caribe (SCALAC), a non-profit civil association that, as of June 2020, brings together nine countries in the region.

Regarding connectivity, the current situation of Latin America is that the main ring of the RedCLARA network (in orange in Figure 4), connecting Brazil, Chile, and Panama with the U.S., has a bandwidth of 100Gbps, and links from 500Mbps to 20Gbps con-

nect the other Latin American countries with the main ring. In turn, the connection with Europe (via France and U.K.) is over 10Gbps, similar to the connection to the U.S. from Mexico.

### International Research Collaborations

Latin America has a long tradition of international collaborations in areas related to e-Infrastructures for science, including grid computing, cloud computing, and HPC. Because of the cultural ties between Latin America and Europe, a large part of these collaborations involved countries in these two regions. Noteworthy is the sequence of projects EELA, EELA-2,

and GISELA, partially funded by the European Commission (EC), whose goals were to deploy and consolidate grid computing infrastructures present in almost all Latin American countries, and connect them to similar European initiatives. Recently, a different initiative from the RICAP project aimed at integrating Latin American e-Infrastructures (both HPC and federated clouds) in a sustainable network. The RICAP project coordinates its activities with SCALAC, agreeing to deploy HPC calls for accessing computing resources via RedCLARA as one of their thematic services. Such initiative has been exercised during the COVID-19 pandemic, when SINAPAD and SCALAC together have made available to researchers prioritized access to HPC and AI resources for pandemic-related projects.

At the national level, Brazil has participated in international collaborations on e-Infrastructure with the BRICS countries, the U.S., and Europe. With the latter, in particular, Brazil co-funded e-Infrastructure projects carried out by large consortia involving research centers and private companies, with MCTI funding the Brazilian partners, while the EC funded the European ones. In the area of cloud computing for science, the EUBrazil Cloud Connect project involved 12 partners to produce technology for the federation of heterogeneous cloud infrastructures in Brazil and Europe, followed by others in this area (EUBra-BIGSEA, SecureCloud, and ATMOSPHERE). In the area of HPC for science, the HPC4e project brought together 13 partners in Brazil and Europe, aiming at going beyond the state-of-the-art in the required HPC simulations for wind energy production and design, efficient combustion systems for biomass-derived fuels, and exploration geophysics for hydrocarbon reservoirs.

Mexico has been involved with advanced computing international collaborations through several projects, for example, at the Large Hadron Collider (CERN), at the HAWC Gamma Ray Observatory (Mexico)<sup>1</sup> and at the Pierre Auger Observatory (Argentina). Recently, the ENERXICO project builds upon the expertise of a consortium of 15 institutions distrib-



uted between Mexico and Europe to deliver groundbreaking new energy solutions, from wind turbine simulations to improve the efficiency of wind farms and make wind energy more competitive, to geophysics exploration and oil reservoir modeling, to thermo- and fluid-dynamic processes of biofuel combustion for transportation. Mexican partners are developing exascale-ready application codes, among them, the first Smoothed Particle Hydrodynamics based code that has ever been elaborated for the numerical simulation of oil reservoirs. Furthermore, to reduce the uncertainties, novel applications of AI are being implemented to mimic the acumen and experience of the experts who need to make an inference based on field characteristics.

Several international collaborations including Uruguayan partners have been developed at the National Supercomputing Center. Some relevant recent projects are: “Geophysics and cosmic rays detection using Hubble Space Telescope (HST)” (National Space Telescope, U.S.), to exploit dark images from HST for cosmic ray detector to analyze Earth’s external magnetic field, applying HPC and cloud computing, to complement measurements from 93 geophysical observatories; Energy efficiency projects with LA/EU partners, ranging from machine learning for data analysis to design of new technologies for renewable energy sources; “Urban transport planning in smart cities” (LA/EU, Mexico/Cuba/Spain/France), studying and implementing methodologies for mobility planning in modern cities, applying computational intelligence and reliable/secure processing of large volumes of data, to assist users and authorities in mobility decision-making (data collection, mobility patterns identification, route/frequencies planning, sustainable mobility, traffic lights synchronization); and the already commented SYRAH-COVID initiative for building a free database of biomolecular simulations of SARS-CoV2 proteins.

### The Way Forward

The current situation of HPC in Latin America is very promising for developing new initiatives and strengthening

existing ones for further growth of science and education.


Regarding infrastructure, the region counts with several platforms and centers that not only provide computing power, but also new-generation services for developing complex research projects. Many of the existing platforms support some degree of interconnection with similar ones in the continent, or even with larger platforms in Europe and the U.S. These systems are focused on high-impact research efforts, especially those related with the current social and economic situation of the countries in this region. It is also expected that by the end of the year 2021, connectivity will be significantly enhanced by the installation of a new submarine cable connecting Europe directly to Latin America at 2Tb/s. This will foster collaboration and allow the development of new capabilities and the access of NRENs to additional HPC resources, using services provided by RedCLARA.

These are exciting times for HPC development, with exascale computational capabilities at the turn of the corner. There is a convergence under way; bringing together HPC and AI, along with data analytics (DA), in what may become the profile of a single, integrated system. What is manifest in recent years is that it is becoming difficult to reach specific research and technological goals without the interplay among these three technologies. Aligned with this trend, LNCC and members of the Mexican Network in HPC including CINVESTAV have recently expanded their infrastructures to provide better support for AI-oriented research.

All the advancement in these new integrated systems will trickle down to society through smaller systems that will incorporate landmark innovations developed in the way to the exascale era and allow researchers to collectively benefit from these new technologies. Latin American public and private centers for research and technology will certainly be using these new systems to solve highly demanding HPC numerical simulations of complex problems, running innovative AI applications and confronting intricate DA challenges. Examples include those involving medical imag-

ing, genomic analysis, astrophysics, climate models, smart cities, and digital agriculture, to mention a few. Moreover, situations like the COVID-19 pandemic calls Latin America for agile collaboration arrangements—such as the one proposed by SCALAC—that cross institutional and governmental boundaries.

### Acknowledgments

This work would not be possible without the help of Artur Ziviani, Augusto Gadelha, Francisco Brasileiro, Pablo Blanco, Jose Maria Cela, Jaime Klapp, Leonardo Sigalotti, Salma Jalife, Rafael Mayo, and the boards of RedCLARA and SCALAC. Map in Figure 1 adapted from original work by Felipe Menegaz under license I CC-BY-SA-3.0 (<http://creativecommons.org/licenses/by-sa/3.0/>). 

### References

1. Abeyssekara, A. et al. Very-high-energy particle acceleration powered by the jets of the microquasar SS 433. *Nature* 562 (2018), 82–85.
2. Arocena, R., Göransson, B., and J. Sutz. Knowledge policies and universities in developing countries: Inclusive development and the developmental university. *Technology in Society* 41 (2015), 10–20.
3. Blanco, P. et al. Comparison of 1D and 3D models for the estimation of fractional flow reserve. *Scientific Reports* 8, 17275 (2018).
4. Fierro-Santillán, C. et al. FITspec: A new algorithm for the automated fit of synthetic stellar spectra for OB stars. *The Astrophysical J. Supplement Series*, 236:38 (2018).
5. Kalinin, N. Self-organized criticality and pattern emergence through the lens of tropical geometry. *PNAS* 115 (35) E8135-E8142 (2018).
6. Klapp, J. et al. Tsunami hydrodynamic force on a building using a SPH real-scale numerical simulation. *Natural Hazards* 100, 1 (2020), 89–109.
7. López-Camacho, C. et al. Rational Zika vaccine design via the modulation of antigen membrane anchors in chimpanzee adenoviral vectors. *Nature Commun.* 9, 2441 (2018).
8. Machado, M. and Pantano, S. Split the charge difference in two! A rule of thumb for adding proper amounts of ions in MD simulations. *J. Chemical Theory and Computation* 16, 3 (2020), 1367–1372.
9. Ocaña, K. et al. BioinfoPortal: A scientific gateway for integrating bioinformatics applications on the Brazilian national high-performance computing network. *Future Generation Computer Systems* 107, (2020), 192–214.
10. Porteiro R., Hernández-Callejo, L., and Nesmachnow, S. Electricity demand forecasting in industrial and residential facilities using ensemble machine learning. *Revista Facultad de Ingeniería, Universidad de Antioquia* (2020).
11. Santos, K. et al. Highly flexible ligand docking: Benchmarking of the DockThor program on the LEADS-PEP protein-peptide data set. *J. Chemical Information and Modeling* 60, 2 (2020), 667–683.

**Isidoro Gitter** is a professor and researcher at the Center for Research and Advanced Studies (CINVESTAV) in Mexico City, Mexico.

**Antônio Tadeu A. Gomes** is a researcher at the National Laboratory for Scientific Computing (LNCC), Petrópolis, Brazil.

**Sergio Nesmachnow** is a professor and researcher at Universidad de la República in Montevideo, Uruguay.



BY MONICA TENTORI, ARTUR ZIVIANI,  
DÉBORA C. MUCHALUAT-SAADE, AND JESUS FAVELA

# Digital Healthcare in Latin America: The Case of Brazil and Mexico

THE HEALTHCARE SYSTEM in Latin America (LATAM) has made significant improvements in the last few decades. Nevertheless, it still faces significant challenges, including poor access to healthcare services, insufficient resources, and inequalities in health that may lead to decreased life expectancy, lower quality of life, and poor economic growth.

Digital Healthcare (DH) enables the convergence of innovative technology with recent advances in neuroscience, medicine, and public healthcare policy.<sup>a</sup> In this article, we discuss key DH efforts that can help address some of the challenges of the healthcare

<sup>a</sup> <https://bit.ly/36l8J6O>

system in LATAM focusing on two countries: Brazil and Mexico.

We chose to study DH in the context of Brazil and Mexico as both countries are good representatives of the situation of the healthcare system in LATAM and face similar challenges along with other LATAM countries. Brazil and Mexico have the largest economies in the region and account for approximately half of the population and geographic territory of LATAM.<sup>11</sup>

Brazil and Mexico each have a government-run public healthcare system that is universal in Brazil and fragmented in Mexico. Most Brazilians and Mexicans depend on the public healthcare system exclusively, and those who can afford it often pay for private health insurance. The demographics of Brazil and Mexico pose significant challenges to health. While life expectancy has experienced a significant increase in recent years, unhealthy lifestyles, poor environmental health, and low adherence to treatments have shifted causes of mortality toward chronic diseases. Thus, Brazil and Mexico are experiencing an increasing demand for uncompensated informal care that is higher than what is reported in developed countries;<sup>15</sup> lower diagnostic and treatment costs are urgently needed.

Research efforts on DH in Brazil and Mexico involve a diversified composition of activities ranging from telehealth to mathematical modeling. Here we discuss several research projects showcasing how DH, using major computing science paradigms, may enhance the efficiency of healthcare delivery in a LATAM context.

## Healthcare Analytics

Although healthcare analytics is a global trend,<sup>14</sup> in LATAM it gains particular relevance given its specificities in demographic and socioeconomic aspects and in the adopted healthcare system. Healthcare analytics enables different projects tailored for the LATAM reality, such as the evaluation of public healthcare policies, big data





A young child undergoing music therapy using Bendable Sound.



## Healthcare analytics, pervasive healthcare, and artificial intelligence solutions have lowered diagnostic and treatment costs to enhance the efficiency of the healthcare delivery in LATAM.

analytics for public health research, or the data-driven analysis of the healthcare system.

An example of an ongoing project in Brazil for the evaluation of public healthcare policies is the assessment of the impact of hospital-based breastfeeding interventions on infant health. Brazil has over three million births per year, 10% of which are premature births.<sup>b</sup> About 25% of the Brazilian hospitals implement at least one of the three initiatives that promote breastfeeding in maternal hospitals. This on-going project,<sup>3</sup> led by the Institute of Scientific and Technological Communication and Information in Health/Fiocruz with the National Laboratory for Scientific Computing (LNCC) and CEFET-RJ as partners, assesses the impact on neonatal mortality of breastfeeding programs (alone or in combination), in particular considering that breastfeeding offers newborns increased protection against infectious diseases.<sup>2</sup> This assessment is performed using spatial and temporal analyses that integrate epidemiological, statistical, and data science tools. To that end, the project analyzes more than 60 million deliveries and about 320,000 neonatal deaths in maternity hospitals in Brazil for over 20 years to better understand the cost-effectiveness of the combined adoption of these public policies promoting breastfeeding. The results thus help define integrated guidelines for the future implementation of these initiatives.

Another ongoing project in Brazil focuses on data-driven analysis and it involves the evaluation of patient trajectories through the healthcare system by using time-varying graphs and process mining tools.<sup>12,13</sup> To that end, this project, involving partner researchers from LNCC, the Federal University of Juiz de Fora, and the Brazilian Ministry of Health, comprises data of almost 60 million healthcare consultations or health procedures delivered to about 6.5 million unique patients in the city of São Paulo over two years (2014–2015). In a case study, the project analyzed the trajectories through the healthcare

system of almost 25,000 pregnant women. The results promote understanding of how the system is being used, its bottlenecks, and where they emerge, thus offering knowledge to build upon for better management and resource allocation.

Research using data-driven analysis and mathematical models to analyze epidemiological data has also been active in Mexico. One ongoing project led by the Institute of Research in Applied Mathematics and Systems at the National Autonomous University of Mexico (UNAM), including collaborations with the Mexican Ministry of Health, has explored the use of time-dependent epidemiological models to make reliable predictions on the evolution of influenza<sup>6</sup> and the COVID-19 pandemics. Results indicate the potential of the sanitary measures in reducing outbreaks.

Overall, these projects contribute to evaluating public healthcare policies to eventually cope with healthcare disparities that are common in LATAM due to an unequal distribution of economic income and assets.

### AI for Healthcare

Artificial intelligence (AI) has shown a growing impact in key categories of healthcare mainly involving the screening and the early detection of different health disorders. Screening is a strategy adopted by health authorities in order to identify a health problem at initial stages and enable early intervention. However, current methods used as the gold standard for screening present multiple limitations including high false positive classification rate or using data that is subject to clinicians' interpretation and often inaccurate and biased. Furthermore, due to having limited equipment in LATAM hospitals, specialized exams for screening including imaging exams, like mammography, magnetic resonance (MR) or computed tomography (CT) are hardly available; so there is a huge need for alternative low-cost exams.

Moving in this direction, a research project, developed by VisualLab at Fluminense Federal University (UFF), proposes thermography image processing using AI techniques for breast cancer screening. Breast cancer is the

<sup>b</sup> [https://www.who.int/pmnch/media/news/2012/201204\\_bornতোosoon-report.pdf](https://www.who.int/pmnch/media/news/2012/201204_bornতোosoon-report.pdf)



leading cause of cancer deaths among women in Brazil and Mexico, reaching almost 30% of cancer deaths in both countries. Since cancerous tissue temperature is generally higher than healthy surrounding tissues, thermography has been considered a promising screening method for breast cancer detection, by generating images that reveal the heat distribution on the skin surface.<sup>9</sup> VisualLab has developed image analysis methods to assess the risk of breast cancer, using dynamic infrared thermography,<sup>16</sup> a method for monitoring the dynamic response of skin temperature after thermal stress. Breast images of UFF's University Hospital patients were analyzed to build time series. Those temperature time series are then clustered by the *k*-means algorithm and evaluated by clustering validation indices. Indices values are treated as features and submitted to a feature selection step. As the last step, classification is performed considering the feature vector for each patient. The method was tested with 1,400 images from 70 different patients (35 healthy and 35 with breast anomalies) acquired with a FLIR SC620 thermal camera. Among several classifiers, the best results were achieved by K-Star, obtaining 98.57% accuracy using tenfold cross-validation.<sup>17</sup>

The Neuroimaging Laboratory at Universidad Autónoma Metropolitana-Iztapalapa, a leading group on image and signal processing for healthcare in Mexico, has also shown how image processing can be used for the screening and monitoring of cognitive decline by identifying anatomical and functional connectivity of the brain from electroencephalogram recordings.<sup>1</sup>

As deep learning makes inroads in medical image processing, being able to explain the results of a machine learning algorithm—a field referred to as explainable AI, or XAI—becomes paramount and has become an important trend for DH systems. For example, an ongoing project, developed by MídiaCom Lab at UFF, involves the use of AI for early detection of patients with mild cognitive impairment (MCI)<sup>4</sup> to promote early interventions and avoid dementia. As patients attending public hospitals in Brazil usually do not have MR or CT imaging



exams, the proposed dynamic decision model uses clinical data, including symptoms and neuropsychological tests. Periodically an ML process is run and evaluates different kinds of AI techniques, combining discretization and balancing methods applied to the available patient dataset, in order to select an updated decision model with the best performance. The model was trained and tested with 319 patients from the Center for Alzheimer's Disease at the Federal University of Rio de Janeiro, and the Center for Attention to Health of the Elderly at UFF, achieving 94% accuracy. In addition to indicating the diagnostic probability, the proposed decision model also informs the physician about which test results have most influenced the system decision and which additional tests can be made to increase diagnostic probability.

Such projects illustrate how ML and image processing can be used for early, low-cost non-invasive breast cancer, Alzheimer's, and MCI screenings in LATAM. The projects from Brazil are part of the research network of the National Institute of Science and Technology in Medicine Assisted by Scientific Computing, covering a diverse set of activities related to DH.

### Pervasive Healthcare


For more than a decade, pervasive healthcare (PH) research in LATAM has focused on the development and pilot-testing of pervasive sensing solutions, natural user interfaces, and

remote care management. Several researchers have proposed mobile and wearable sensing platforms, involving either user or environment instrumentation, to collect patient data that is relevant for clinical case assessment and could be used to train machine learning models for the early detection and diagnosis of different disorders and diseases. A recent trend is to use digital markers to quantify physiological and behavioral data collected through wearable and mobile sensing, or through the tracking of interaction patterns with digital devices.


One example is an ongoing project from a Mexican autism network that has deployed the first living laboratory<sup>10</sup> actively using PH technology in the everyday activities of a specialized school-clinic. The network is currently being led by the Ensenada Center for Scientific Research and Higher Education (CICESE) and includes academic institutions like Centro de Enseñanza Técnica Y Superior University and UNAM, and different public and private schools located in northwest Mexico. This project explores if elastic displays can uncover force-related gestural interactions that can be used as a digital marker to support autism screening.

The elastic display BendableSound, developed at the Mobile and Ubiquitous Computing Lab at CICESE-Mexico, runs a 3D background of space-based elements that play sounds when users touch them.<sup>5</sup> A study of children with autism and neurotypicals who





**Socially assistive robots are interactive, intelligent systems that employ hands-off social interaction strategies, including the use of speech, facial expressions, and communicative gestures, to assist in a particular healthcare context.**



used BendableSound uncovered atypicalities in the amount of force and gestural interactions used by children with autism. Machine learning models using the more dominant features and the more relevant gestures show good performance in classifying children with autism from neurotypicals with 97.3% of precision and recall.

Other research initiatives in LATAM have proposed the use of telemedicine, social media, and crowdsourcing for remote care management, as Brazil and Mexico have large territories with remote rural and unplanned metropolitan areas that are hard to reach. This research shows how PH can be used to improve patient-clinician communication and support remote patient monitoring. Telehealth consultations, whose regulation faced a long history in Brazil and Mexico, was suddenly regulated in Brazil and was being offered to those socially isolated at home during the pandemic of 2020.<sup>c</sup>

PH research in LATAM has also been actively studying the use of natural user interfaces to support therapeutic interventions. Solutions range from the use of brain-computer interfaces to the use of gesture therapy,<sup>18</sup> a term coined at the Robotics Lab at the National Institute for Astrophysics, Optics, and Electronics to reduce the gap between cost and accessibility, lowering therapeutic costs and lessening the burden of informal caregivers in LATAM.

A success story is the use of socially assistive robots (SARs) to assist in the care of older adults with dementia. SARs are interactive, intelligent systems that employ hands-off social interaction strategies, including the use of speech, facial expressions, and communicative gestures, to assist in a particular healthcare context.

The robot Eva, developed at the Mobile and Ubiquitous Computing Lab at CICESE-Mexico, is an open source SAR using conversational strategies for people with dementia as proposed by organizations such as the Alzheimer Association.<sup>7</sup> Eva relies on knowledge about the user, including their capabilities and preferences, to personalize and guide cognitive stimulation therapy using activities such as music, remi-

niscence, cognitive games, and relaxation. It incorporates new knowledge from the user after each session of use. While following a somewhat limited, but personalized script, the robot can autonomously lead a therapy session with minimal conversation breakdowns. The robot has been evaluated with individuals with mild to moderate dementia for the last three years in a geriatric residence. The results show older adults with dementia learn to interact with Eva with no previous training in one or two sessions. After nine weeks of using Eva, older adults significantly decreased their neuropsychiatric symptoms associated with dementia and increased their quality of life.<sup>8</sup> Some of the technical challenges facing the widespread use of social robots for dementia include developing proper adaptation strategies based on disease progression and inferring user mood. In addition, the continuous use of a conversational agent could be used to assess the course of the disease and the effectiveness of interventions through paralinguistic or verbal analysis.

These efforts represent projects showing PH can make digital healthcare solutions increasingly available in the homes, geriatric residences, and in the school-clinics in LATAM.

### **Challenges and Trends**

DH research is representative in Mexico and Brazil due to the particularities of their demographics, culture, economic, and geographic territory. In this article, we showed how healthcare analytics, AI, and PH may improve their healthcare system. Here we discuss major challenges as a springboard for a new focus on DH in LATAM.

Health data collection is very challenging and expensive. The LATAM context imposes significant shortcomings to collect and maintain large public datasets with consistent and reliable information. Current key challenges involve the homogenization of diverse levels of data quality and accuracy as well as the integration of data fragments from diverse sources. The key to showing the current value of such large datasets involves the challenging process of converting big data into valuable insights which may

<sup>c</sup> <https://perma.cc/EF7X-Y3PX>

be limited due to the low number of cases, class imbalance, and nonuniform misclassification in current AI models. Risks associated with the tensions of collecting data in the lab versus in naturalistic conditions have resulted in solutions divorced from practical applications when working with real-world scenarios resulting in poor performance due to using unreliable data.

The feasibility of the collection of health data during everyday activities will only be possible through the use of innovative technology that individuals are using on a daily basis. This opens up opportunities for the creation of natural and intuitive forms of interaction and innovative devices to ease data gathering and the analysis of large datasets in naturalistic conditions. Development challenges involve exploring novel input and output mechanisms through custom-built wearable and sensing devices, the use of natural user interfaces to measure novel gestural interactions, and techniques to indirectly infer health data. In LATAM, healthcare clinics and hospitals have limited equipment, making it difficult to deploy innovative technology without an appropriate information technology backbone; understanding what infrastructure and architectures are needed to use built-in sensors and innovative technology will help healthcare providers deal with troubles of upscaling and scalability.


Moreover, general DH solutions have not been developed to treat the diverse healthcare needs of Latin American citizens. Most DH solutions have typically offered a “one-size-fits-all” solution in which personalization of healthcare services to the particular needs and capabilities of its patients has been largely neglected. A current trend in AI involves the development of predictive and adaptive decision models to be used in different healthcare scenarios. Applying a particular model to a different set of data will not be able to adapt itself; however, one of the main drawbacks of these models is that they do not take into account the context of its use and of that of their users.

Descriptions of the development and empirical study of DH in concrete

scenarios are scarce and urgently needed. The actual deployment of DH in the healthcare system infrastructure poses a variety of social, cultural, legal, and policy-based issues. The area’s average investment in healthcare is below that of developed countries.<sup>d</sup> In particular, clinical and computing methods for the development and evaluation of the impact of DH solutions in concrete scenarios are also different. The low number of participants enrolled in pilot studies and the lack of replication studies have also reduced the possibilities of convincing clinicians and the pharmaceutical industry to invest in transforming DH solutions that come out of research projects into commercial products, and are also limiting research in healthcare analytics and AI.

At the time of writing, Brazil and Mexico, like the rest of the world, were facing the outbreak of the novel coronavirus (SARS-CoV-2). At the center of the resulting health, economic, and social crisis, an urgent call for actions on the application of information and communication technologies to healthcare has emerged, in addition to the more obvious demands for research on epidemiological surveillance and models, drug and vaccine development, and multi-omics analysis. Some initiatives are investigating the feasibility of detecting COVID-19 through AI-based lung image analysis or how AI may accelerate the process of *in silico* drug design. These are just some examples out of many possible ones. Despite the hard outcomes, this crisis has put computing applied to healthcare in the front line, overcoming a tipping point and boosting DH in Brazil and Mexico, with innovative technology playing a key role in enabling better and low-cost healthcare services and assistance.

### Acknowledgments

M. Tentori thanks the support of CONACYT and MSR. A. Ziviani thanks the support of CNPq and FAPERJ. D.C. Muchaluat-Saade thanks the support of CNPq, FAPERJ, INCT-MACC and CAPES PRINT. J. Favela thanks the support of CONACYT. 

<sup>d</sup> <https://www.oecd.org/els/health-systems/health-expenditure.htm>

### References

1. Barabá-Morales E. et al. Evaluation of brain tortuosity measurement for the automatic multimodal classification of subjects with Alzheimer’s disease. *Comput. Intell. Neurosci.* (2020).
2. Boccolini, C.S., Carvalho, M.L., Oliveira, M.I.C. Factors associated with exclusive breastfeeding in the first six months of life in Brazil: a systematic review. *Revista de Saúde Pública*, 49, 91 (2015).
3. Boccolini, C.S. Assessing the impact of hospital-based breastfeeding interventions on infant health, 2020; DOI: 10.7303/syn18428446.
4. Carvalho, C. M. et al. A clinical decision support system for aiding diagnosis of Alzheimer’s disease and related disorders in mobile devices. In *Proceedings of the 2017 IEEE Intern. Conf. Communications*.
5. Cibrian, F.L. et al. BendableSound: An elastic multisensory surface using touch-based interactions to assist children with severe autism during music therapy. *JHCS* 107 (2017), 22–37.
6. Cruz-Pacheco G. et al. Modelling of the influenza A(H1N1) outbreak in Mexico City, April-May 2009, with control sanitary measures. *Euro Surveill*. 14, 26 (2009), 19254.
7. Cruz-Sandoval, D., Favela, J. A conversational robot to conduct therapeutic interventions for dementia. *IEEE Pervasive Computing* 18, 2 (2019), 10–19.
8. Cruz-Sandoval et al. (2020). A social robot as therapy facilitator in interventions to deal with dementia-related behavioral symptoms. In *Proceedings of the 2020 ACM/IEEE Intern. Conf. Human-Robot Interaction*, (Mar. 2020) 161–169.
9. EtehadTavakol, M. et al. Breast cancer detection from thermal images using bispectral invariant features. *Intern. J. Thermal Sciences* 69 (2013), 21–36.
10. Favela, J. et al. Living labs for pervasive healthcare research. *IEEE Pervasive Computing* 14, 2 (Apr.-June 2015), 86–89.
11. Horwitz, B., Bagley, B.M. *Latin America and the Caribbean in the Global Context. Why care about the Americas?* Taylor and Francis Inc., 2016.
12. Miranda, M.A. et al. Characterization of the flow of patients in a hospital from complex networks. *Health Care Management Science* 23 (2020), 66–79.
13. Rebugea, Á., Ferreira, D.R. Business process analysis in healthcare environments: A methodology based on process mining. *Information Systems* 37, 2 (2012), 99–116.
14. Science Europe—Medical Sciences Committee. How to transform big data into better health: Envisioning a health big data ecosystem for advancing biomedical research and improving health outcomes in Europe. Workshop Report, Erice, Italy, 2014.
15. Shahly, V. et al. Cross-national differences in the prevalence and correlates of burden among older family caregivers in the World Health Organization World Mental Health (WMH) Surveys. *Psychological Medicine* 43, 4 (2013), 865–879.
16. Silva, L.F. et al. Hybrid analysis for indicating patients with breast cancer using temperature time series. *Computer Methods and Programs in Biomedicine* 130 (2016), 142–153.
17. Silva L.F. et al. A computational method for breast abnormality detection using thermographs. In *Proceedings of the IEEE 33rd Intern. Symp. on Computer Based Medical Systems*, 2020.
18. Suca, L.E. et al. Gesture therapy: A clinical evaluation. *PervasiveHealth*, 2009.

**Monica Tentori** is an associate professor in the Department of Computer Science at the Ensenada Center for Scientific Research and Higher Education, Ensenada, Mexico.

**Artur Ziviani** is a senior researcher at the Data Extreme Lab of the National Laboratory for Scientific Computing, Petrópolis, Brazil.

**Débora C. Muchaluat-Saade** is a professor in the Institute of Computing and the head of MídiaCom Lab at Fluminense Federal University, Niterói, Brazil.

**Jesus Favela** is a professor in the Department of Computer Science at the Ensenada Center for Scientific Research and Higher Education, Ensenada, Mexico.



BY MARCELO ARENAS AND PABLO BARCELÓ

# Chile's New Interdisciplinary Institute for Foundational Research on Data

THE MILLENNIUM INSTITUTE for Foundational Research on Data<sup>a</sup> (IMFD) started its operations in June 2018, funded by the Millennium Science Initiative of the Chilean National Agency of Research and Development.<sup>b</sup> IMFD is a joint initiative led by Universidad de Chile and Universidad Católica de Chile, with the participation of five other Chilean universities: Universidad de Concepción, Universidad de Talca, Universidad Técnica Federico Santa María, Universidad Diego Portales, and Universidad Adolfo Ibáñez. IMFD aims to be a reference center in Latin America related to state-of-the-art research on the foundational problems with data, as well as its

a <https://imfd.cl/en/>

b [http://www.iniciativamilenio.cl/en/home\\_en](http://www.iniciativamilenio.cl/en/home_en)

applications to tackling diverse issues ranging from scientific challenges to complex social problems.

As tasks of this kind are interdisciplinary by nature, IMFD gathers a large number of researchers in several areas that include traditional computer science areas such as data management, Web science, algorithms and data structures, privacy and verification, information retrieval, data mining, machine learning, and knowledge representation, as well as some areas from other fields, including statistics, political science, and communication studies. IMFD currently hosts 36 researchers, seven postdoctoral fellows, and more than 100 students.

We at IMFD are convinced the development of a science of data requires producing a virtuous amalgamation of all the aforementioned areas, in order to cope with ever-increasing societal demands for taking full advantage of available information. A dramatic example of such needs has been given to humanity by the current COVID-19 pandemic, but there are clearly many others. At IMFD, we are carrying out several projects that aim to produce such interdisciplinary crossings.

More importantly, we have worked to develop a common research agenda for the Institute, integrating the efforts of its members into interdisciplinary teams whose goal is solving some fundamental problems in data science. Specifically, the research agenda of IMFD has been organized into five long-term and transversal *emblematic* research projects, each of which requires input from several, if not all, the research areas mentioned previously. These emblematic projects are:

► *Data for the study of complex social problems.* This project seeks to encourage the combination of methodological strategies and techniques based on data analysis to develop novel diagnoses about relevant socio-political issues.









Members of the Millennium Institute for Foundational Research on Data at a workshop held on Viña del Mar, Chile, in 2019.

► *Development of systems for graph and network analysis.* This project focuses on developing an efficient, correct implementation of standard languages to extract information from graph-structured data, as a way to retrieve valuable analytic information from large networks of interconnected data.

► *Query answering methods for emerging requirements.* This project seeks to develop theoretical foundations for information extraction tasks capable of integrating data management, data analysis, and machine learning techniques, while also making them scalable and verifiable.

► *Explainable artificial intelligence.* This project aims to develop new techniques that allow understanding of the inference processes of machine learning algorithms, where this understanding is embodied in the ability to provide explanations for learned patterns, as well as to create higher-level abstractions and procedures based on them.

► *Development of robust information structures.* This project pursues understanding of whether it is possible to create systems that allow people to have a more-accurate, less-biased view of reality, and whether the

volume of data generated by users in digital platforms can be used to obtain a clearer picture of the social uses and communication phenomena of the Web.

The goal of this article is to show the achievements of IMFD along these lines of research, as well as to briefly reflect on the future of the Institute.

**Some Achievements of the Institute Data for the study of complex social problems.**

To meet the goal of this project, we have been developing a methodology for gathering and maintaining thick data,<sup>c</sup> which is data gathered by using qualitative methods, and for combining such data with large online sources to study the socio-political and economical dynamics within Chilean territory. Such thick data is usually collected from surveys, so we have focused on four territorial enclaves in contemporary Chile that concentrate a host of sociopolitical and socioeconomic challenges related to crucial dimensions of social life. Moreover, we have worked to produce prospective scenarios for public policy-making in

each territorial enclave through computational social science techniques like agent-based modeling, drawing key parameters from the four territorial enclaves.

We currently are working on several lines of research in this project. In what follows, we explain one of them to exemplify the methodology mentioned earlier. In 2019, we designed and launched the “Monitor” project, which aims to produce a white paper each year on specific public policy challenges and their manifestations/implications in/for each type of territorial enclave. Our first coordinated fieldwork was conducted in the Quinteros bay area, one of the zones of environmental sacrifice in Chile. A sacrifice zone is a concept that emerged in the U.S. during Nixon’s presidency as a result of the installation of coal and nuclear power plants in Utah, New Mexico, Arizona, and Colorado. Today, a sacrifice zone is a term used by different scholars to characterize specific places where the population (generally poor) coexists with industries whose activities are mainly based on coal, oil, gas, or nuclear energy. The aim of this fieldwork was twofold. On the one hand, we conducted in-depth interviews

<sup>c</sup> <https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7>



aimed at understanding how socio-political dynamics are carried out in these zones, identifying key challenges for public policies, and understanding why these challenges are not addressed. On the other hand, it served us as a way of producing more input to address some of the research questions we are working on, such as how environmental sacrifice areas are created. Faithful to our thick-data methodology, we combined this information with a large database of Chilean news about the area, in order to understand public opinion about zones of sacrifice. The results of this combination were promising; in particular, we have learned many lessons from them on how to combine thick data with online sources.

**Development of systems for graph and network analysis.** Graph databases are a fast-growing technology for modeling data and extracting information, in particular, because of the large number of applications where data can be naturally represented as a graph (as examples, consider social, crime detection, scientific, and bibliographic networks). This has created great interest from academia and industry in standardizing languages for extracting information from graph databases. The research background of our group, including several foundational papers in the area,<sup>6</sup> as well as our participation in standardization processes, gives us a comparative advantage to become key actors in the future of graph database theory and practice.

The main goal of this project is to develop languages to extract information from graph databases. In particular, we aim to develop a full-scale graph database system that encompasses state-of-the-art techniques for all aspects of data management, from data storage and indexing to concurrent access, querying, and graph analytics, and which can deal with large volumes of data. To this end, we have consolidated several lines of research, two of which will be discussed later.

*Development of a standard graph query language.* We actively led an effort between industry and academia to develop a standard graph database query language. As a result of this ef-

fort, we proposed the query language G-CORE,<sup>1</sup> which fulfills three fundamental principles for graphing such languages: to have a careful balance between expressiveness and evaluation complexity; to be composable (meaning that graphs are both the input and the output of queries), and to treat paths as first-class citizens. We are convinced that G-CORE will play a key role in shaping the future of graph query languages. In fact, the International Organization for Standardization (ISO) has launched an initiative to standardize graph query languages; two members of IMFD are participating, and G-CORE is considered one of the role models.

*Creation of efficient algorithms for query answering.* Joins are the costliest operation in query processing, and they have been the subject of intense research. Recent studies show that joins can be handled optimally by using appropriate data structures, which unfortunately are computationally hard to build.<sup>2</sup> One of the central lines of research of this project is the development of new techniques to handle joins within compact space and with provable performance guarantees. In particular, we have developed an algorithm to process join queries over a compressed representation of graphs and show that its running time is worst-case optimal,<sup>3</sup> and we have shown how worst-case optimal join algorithms can significantly speed up the performance of the graph query language SPARQL,<sup>4</sup> the standard query language for Semantic Web data. Moreover, we have explored new paradigms for answering queries over large volumes of data; in particular, we have studied the problems of enumerating, uniformly generating, and counting the answers to a query, proposing a simple yet general unifying framework to investigate these fundamental algorithmic problems, in particular for the case of graph databases.<sup>5</sup>

**Query answering methods for emerging requirements.** As the requirements for data management systems continue to evolve at an accelerating rate, the diversity of proposed solutions addressing these requirements likewise continues to grow.



**IMFD aims to be a reference center in Latin America related to state-of-the-art research on the foundational problems with data, as well as its applications to tackling diverse issues ranging from scientific challenges to complex social problems.**







**We carried out the first empirical study of public opinion on false news in Chile. This work outlines who spreads, and how they spread, false news in the country.**



While technology is rapidly advancing in sub-areas such as databases, machine learning, data analytics, information retrieval, privacy, and so on, the techniques developed are becoming increasingly specialized and divergent from each other. This project aims to help unify those currently disparate techniques by looking into several specific directions.

The ultimate goal of our project is to provide theoretical grounds for the next generation of database systems, trying to make them more flexible, scalable, secure, and robust. We have started, however, by pursuing some objectives that are more at-hand, and that can be seen as the first steps toward our more ambitious goals. These first steps have been achieved by bringing together researchers with different expertise in areas relevant to the project and making them work in specific projects that contribute toward a particular unexplored aspect of data management and its relationship with neighboring fields. Among others, we have started working on developing formal methods for verifying the semantics of query languages in systems like Coq;<sup>20</sup> building and studying languages for expressing analytical queries, like languages that combine features from relational and linear algebra;<sup>7</sup> characterizing the expressive and computational power of modern neural network architectures, including Transformers and Neural GPUs,<sup>13</sup> as well as Graph Neural Networks;<sup>8</sup> formalizing the notion of in-database classification of entities by developing and studying a framework that classifies entities based on features defined as relational database queries;<sup>9</sup> and, finally, building efficient algorithms for evaluating complex analytical queries over streams of data.<sup>10,11,12</sup>

An important stepping stone toward the full realization of the objectives of this project was the organization of an international workshop, called Emerging Challenges in Databases and AI Research (DBAI), in Santa Cruz, Chile, during November 2019.<sup>d</sup> The main objective of the workshop was to bring together a number of different communities to work on several of the problems mentioned here but

<sup>d</sup> <http://dbai2019.imfd.cl/>

from different angles, and to discuss ways in which all this work could be assembled toward the construction of more robust data management systems.

**Explainable artificial intelligence.**

This project focuses its efforts on the development of new techniques that allow understanding of the inference processes behind some artificial intelligence (AI) algorithms, where this understanding is embodied in the ability to provide an explanation for such processes. Since the launching of our Institute, we have consolidated our progress into three main lines of research:

- ▶ *Visual Query Answering*, by developing an AI system capable of applying natural language to explain the reasoning behind the answer to a visual question.<sup>14</sup> This has been enhanced recently through the incorporation of a common-sense knowledge base, with promising results.<sup>15</sup>

- ▶ *Visualization*, by exploring the intersection between information visualization and explainable AI techniques; in particular, the effect of different types of explanations on the reliability of AI systems, both in recommender systems<sup>16</sup> and in document classification systems. Our work in visualization is receiving special attention for medical applications.

- ▶ *Social media*, by focusing on the extraction of information from social media, with the aim of providing explanations of when and why polarizing and controversial effects occur. In the work of this project, early detection of users' stances, harassment detection, early fake news detection, and the study of polarization dynamics in social media occupy central roles.<sup>17,18</sup>

**Development of robust information structures.**

This project focuses on two prominent information disorders; that is, problems that work against the development of robust information structures. These disorders are the spread of misinformation (for example, rumors, conspiracies, hoaxes, and unverified news) and fabricated content (for example, "fake news," propaganda) on online social networks, and hate speech and incivility on digital media.

*The spread of misinformation.*

Informed by social scientific theories,

computational methods, surveys, and quantitative content analysis techniques, this research group continued examining the problem of the spread of incorrect information on social media platforms with several projects that bring together computer scientists, communication scholars, and political scientists. We carried out the first empirical study of public opinion on false news in Chile.<sup>19</sup> This work outlines who and how they spread false news in the country, among other issues related to disinformation. Also, via a project funded by The Social Science Research Council (SSRC) on fake news on Facebook during elections, we have started studying elections in three countries in Latin America: Chile, Colombia, and Mexico. This project aims to characterize the scope and diffusion of fake news in Spanish-speaking countries with respect to verified news, and to delineate the threat of digital disinformation in the region. Finally, we launched a project that takes a social scientific approach to the study of exposure, beliefs, and sharing of conspiracies and fake news in Chile during the social unrest that started in October 2019. Using longitudinal public opinion surveys, this project covered the social uprising and protests to study the sociodemographic, psychological, and media orientations that predict exposure to false information on social media.

#### *Promoting healthy civic conversations on online social networks.*

Through an interdisciplinary perspective, we are studying how conversations take place on social media. Research into the area of incivility and social networks is organized around three projects. The first one analyzes incivility in commentaries posted by users on news media websites. The second project aims to train a classifier to tag comments and classify them as civil/uncivil, within a certain margin of error. The third project studies the relationship between the use of political memes and the presence of incivility in Chilean Twitter accounts.

#### **The Future**

We mentioned in the introduction the current COVID-19 pandemic as a dramatic example of the role that

data is taking in contemporary society. Besides the classical tasks such as development of algorithms and tools to analyze data, deep debates have arisen on fundamental questions about data, such as its public availability; ownership; auditability of the processes used to collect, curate, integrate, analyze, and publish data; balance between transparency and privacy, and the roles that states, universities, research institutions, private organizations, and the general public should play in such issues.


We are convinced IMFD must play a key role in these discussions at the Latin American regional level. In particular, scientific discoveries and technological advances at IMFD must be placed at the service of the development of an integrated Chilean data infrastructure and governance. Hence, in the coming years, we will work on such development, paying special attention to the following five issues: to improve the ways in which data is collected by the Chilean government; to provide unified, integrated access to the data collected and the information developed from it, which should include data curation, but at the same time be auditable; to define different degrees of access to such information, taking into account the tension between transparency and privacy, as well as the different local uses of data; to improve the ways in which such information is analyzed, considering that such processes should be transparent and auditable; and to make this infrastructure one of the first places where the algorithms and techniques developed in IMFD are applied.

The virtual world, and its most basic support, data, came to the forefront with the COVID-19 crisis. This new world is the goal of IMFD research in the immediate future. We are dedicated to helping build the global data governance system (that is, the technical and legal infrastructure), and the social, political, and ethical practices to be observed in the virtual world. Our Institute will follow closely how this new reality is transforming the material practices of areas such as health and education, human life research, digital automation of work, and environmental research. We will focus on the development of areas that

make the virtual world of data a contribution to improve people's lives.

Please join us on this ambitious project, IMFD is an open environment for collaboration!

#### **Acknowledgments**

We thank Claudio Gutiérrez and Sergio Toro for their many useful comments on this document. 

#### **References**

- Angles, R., et al. G-CORE: A core for future graph query languages. In *Proceedings of SIGMOD Conf.*, 2018, 1421–1432.
- Hung Q. Ngo, H.Q., Ré, C., and Rudra, A. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec.* 42, 4 (2013), 5–16.
- Navarro, G., Reutter, J.L., and Rojas-Ledesma, J. Optimal joins using compact data structures. *ICDT*, 2020, 21:1–21:21.
- Hogan, A., Riveros, C., Rojas, C., and Soto, A. A worst-case optimal join algorithm for SPARQL. *ISWC I* (2019), 258–275.
- Arenas, M., Croquevielle, L.A., Jayaram, R., and Riveros, C. Efficient logspace classes for enumeration, counting, and uniform generation. *ACM PODS*, 2019, 59–73.
- Angles, R., et al. Foundations of Modern Query Languages for Graph Databases. *ACM Comput. Surv.* 50(5): 68:1–68:40 (2017).
- Barceló, P., Higuera, N., Pérez, J., and Subercaseaux, B. On the expressiveness of LARA: A unified language for linear and relational algebra. *ICDT*, 2020, 6:1–6:20.
- Barceló, P. et al. The logical expressiveness of graph neural networks. *ICLR*, 2020.
- Barceló, P., Baumgartner, A., Dalmau, V., and Kimelfeld, B. Regularizing conjunctive features for classification. *ACM PODS*, 2019, 2–16.
- Grez, A., Riveros, C., and Ugarte, M. A formal framework for complex event processing. *ICDT*, 2019, 5:1–5:18.
- Grez, A., and Riveros, C. Towards streaming evaluation of queries with correlation in complex event processing. *ICDT*, 2020, 14:1–14:17.
- Grez, A., Riveros, C., Ugarte, M., and Vansummeren, S. On the expressiveness of languages for complex event recognition. *ICDT*, 2020, 15:1–15:17.
- Pérez, J., Marinkovic, J., and Barceló, P. On the Turing completeness of modern neural network architectures. *ICLR*, 2019.
- Zhang, Y., Niebles, J.C., and Soto, A. Interpretable visual question answering by visual grounding from attention supervision mining. In *Proceedings of IEEE Winter Conf. Applications of Computer Vision* (2019).
- Lobel, H., Vidal, R., and Soto, A. CompactNets: Compact hierarchical compositional networks for visual recognition. *Comput. Vis. Image Underst.* 191, 102841 (2020).
- Dominguez, V., Messina, P., Donoso-Guzmán, I., and Parra, D. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of Intern. Conf. Intelligent User Interfaces* (2019).
- Bugueño, M., and Mendoza, M. Applying self-attention for stance classification. *CIARP*, 2019, 51–61.
- Bugueño, M., and Mendoza, M. Learning to detect online harassment on Twitter with the transformer. In *Proceedings of PKDD/ECML Workshops* (2019), 298–306.
- Valenzuela, S., Halpern, D., Katz, J.E., and Miranda, J.P. The paradox of participation versus misinformation: social media, political engagement, and the spread of misinformation. *Digital Journalism* 7, 6 (2019), 802–823.
- Diaz, T., Olmedo, F., and Taner, E. A mechanized formalization of GraphQL. *CPP*, 2020, 201–214.

**Marcelo Arenas** is a professor at the Universidad Católica and the Director of IMFD in Santiago, Chile.

**Pablo Barceló** is a professor at the Universidad Católica and the Deputy Director of IMFD in Santiago, Chile.

© 2020 ACM 0001-0782/20/11



BY GERARDO TORRES ZELAYA

# A Panorama of Computing in Central America and the Caribbean

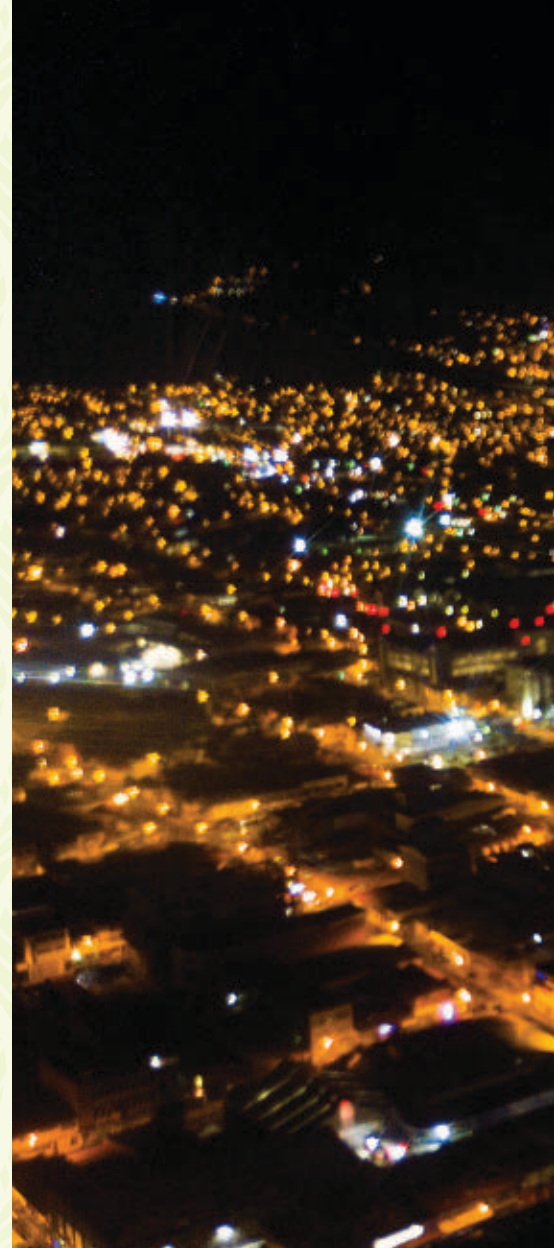
DESPITE BEING A poor and unequal country, Costa Rica has managed to close the gap in access to technology for its citizens, and it is now leading the way in the region. The country started the process of admission for the Organization for Economic Cooperation and Development (OECD) several years ago with reforms on laws, the creation of policies and the use of Computer Technologies to improve education, information access, financial markets, competitiveness, and a more open government. In May 2020, Costa Rica became the first Central American or Caribbean country invited to become an OECD member.

The OECD has almost 60 years of existence, and its members are many of the world's more developed countries that work together to shape policies that

foster prosperity, equality, opportunity, and well-being for their citizens. Costa Rica will become the 38<sup>th</sup> member, the fourth of Latin America.

Education and public investment, linked by the political career of a father and son, are part of the background that explains how Costa Rica's accomplishment was possible in the Central America and Caribbean region's complicated landscape for the development of computer technologies.

Central America and the Caribbean countries are the poorest of Latin America, which, at the same time, is the most unequal region in the world. "Costa Rica bets on education" is a common saying in the region, and dates back to the founding of the Second Republic in 1948, which after a civil war resulted in the abolition of the Armed Forces and a very strong







investment in public education. Fifty years later, the son of one of the main actors in that event (the political leader José Figueres), President José María Figueres inaugurated in 1998 the first Intel factory in Central America that quickly began to assemble 33% of the microprocessors of the Celeron line.

### **Leading the Way**

At that same time (1998) in the rest of the countries of Central America and the Caribbean the economy was still highly dependent on the production and export of fruits, and the clothing manufacturing industry was barely accommodating, when Costa Rica made the leap. Investment in technology, infrastructure, and highly skilled labor (thanks to decades of support for education) attracted companies like Hewlett-Packard (HP) and others to Costa Rica. The government

decided that universal access to the Internet was the way to continue, and now, more than 20 years later, the harvest of that investment has become more evident.

Costa Rica's National Learning Institute and the Technical College offer permanent computer courses for free, and people register aiming to find a job in the country's technology industry. There are also many courses for computer programming, and the development of apps and software. There are bilingual courses such as Information Technological Support, and Computer Networking, that every day become more popular.

The high quality of public and private education systems is one of the country's main offers to attract international investment. The most important in recent years are Consumer Electronics (Noxtak, Matthews

International), Contract Manufacturers (Zollner Electronics, ClamClea), Electronic Components (Electrotechnik) Engineering, Design Camp Software (INTEL Megalab, INTEL National Instruments), Digital Services (Catalina, Cheetah Digital), Digital Technologies (GBT Technologies, Microsoft, NTT Data Inc.) and Engineering Camp (Design: NCI Building Group, Smile Direct).<sup>4</sup>

Technology exportation is mainly focused to the U.S. and even though since 2011 they signed a Trade Agreement with China, there is still not a significant increase in their market exchange with that country, nor presence of Chinese investment in Costa Rica or transcendent commercial relations with Asia. This clear site opportunity is conditioned by political tensions from internal and external groups that have even called to cancel the agreement.



San Jose, the country's capital, was the host of the 43<sup>rd</sup> meeting of the Internet Corporation of Assigned Names and Numbers (ICANN 43) in 2012; and the country has reaffirmed its position in favor of a multi-stakeholder model of Internet governance in the International Telecommunications Union (ITU), and World Conference on International Telecommunications (WCIT) meeting in Dubai during 2012; the Internet Governance Forums (IGF) meetings in 2013–2016; and recent meetings of the Freedom Online Coalition (FOC).

The Costa Rican National Telecommunications Development Plan (PNDT) 2015–2021 for populations in vulnerable conditions includes “Connected Communities” with a budget of \$168 million, aimed to provide Internet services to public education and health institutions in 184 districts throughout Costa Rica. “Connected Homes” with a budget of \$100 million, for subsidized Internet access to 140,000 households in states of poverty and extreme poverty (some 507,000 people), covering about 46% of Costa Rican households in state of poverty and extreme poverty. “Equipping Public Centers” with a budget of \$20 million, is a program that provides computers to public centers for people with disabilities, children, youth, elderly, indigenous people, and women who are heads of household and micro-entrepreneurs. “Connected Public Spaces” has a \$10 million budget to connect 240 public access points (hotspots) with free In-

ternet around the country. “Solidarity Broadband Network” targets public service centers that have higher connectivity needs and aims to improve the speed/quality of their services. “ICT Population Empowerment” program improves digital literacy and focuses on education and online security. “National Teachers Training Program on ICTs” is a program to train teachers on using ICTs in the classroom and in return, teaching their students ICT use. “Technology Platform” program is designed to increase the use of digital technologies by children and young people at 317 schools of the Ministry of Public Education.<sup>10</sup>

### Technology in Difficult Terrain

Central America has a considerably smaller market and development of computer technology than the rest of Latin America. According to the OECD, the biggest investments are concentrated in Costa Rica (manufacture of computer parts, representation of important worldwide companies), Panama (representation of important worldwide companies) and Guatemala (creation of new successful software companies).

The lack of development of computer technologies in the Central American countries is due to their economic situation, given that few companies have the capacity or are willing to invest in such technologies. Even though many universities have started with mechatronic degrees and have created alliances with their

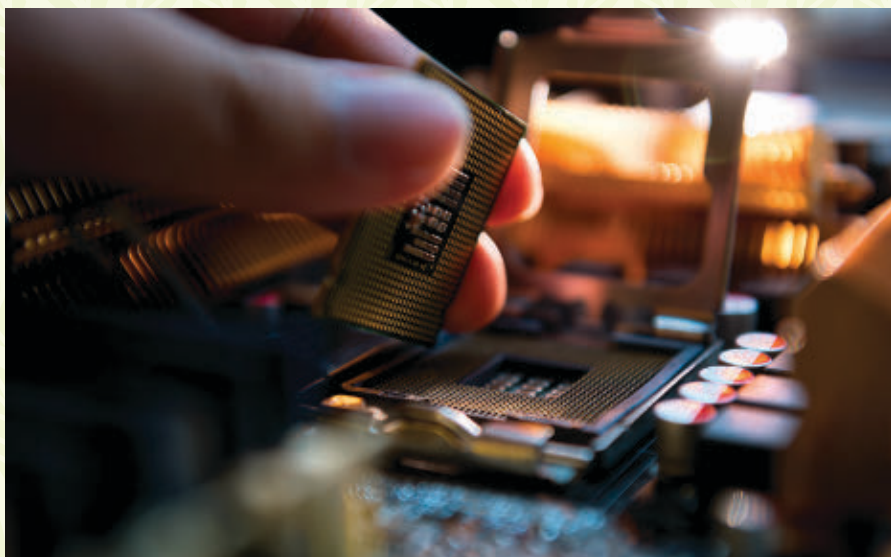
governments to begin the preparation of more qualified employees that can attract international investment, and considering that the World Bank and the International Development Bank are pushing for programs that can help move the existing clothing factories' infrastructure and begin to create industrial parks for the manufacture of computer parts or the development of technology companies outside Costa Rica, the region still does not attract much international industrial technology investment.

There are many small investments focused on app development and computer technology assistance for governments or international programs in the region. Many new companies have been created mainly by young engineers that studied in the several universities in the region that started offering many bachelor's and master's degrees on computer programming, computer systems, mechatronics, and robotics 15 or 20 years ago. Many of them earned their bachelor's degree in their homeland and then traveled to the U.S. for master's and Ph.D's.<sup>5</sup>

Computer technology development has been related more to academia and research than to productivity or investment. The region's first computer center was created in Cuba in the early 1960s and then Costa Rica had also the use of computers in the mid 1960s in the National University. Both countries have maintained their tradition and currently offer strong Doctorates in computer science, specially Cuba that this year in April organized a World Forum on Computer Technology with the intention to strengthen the relationship with China and Latin American Universities and Scientific Centers.<sup>11</sup>

Puerto Rico, Dominican Republic, and the Bahamas also are pushing strong in their Ph.D. programs, and now almost all the countries have master's degrees for programming, TICs, or computer science. The problem remains that the economic sector does not have the capacity to correctly absorb all this high skilled workforce.

Most graduates get certified in the use of programs, because in the search for clients or a job, certifications are more profitable than masters. “Coursera” is one of the many platforms that



young engineers and programmers are using to certify themselves.

Another option is entrepreneurship. According to the Manifest count down of the top 100 mobile app developers in Latin America (May 2020) Central America has 10 companies in that list: Gorilla Logic (Costa Rica), Hatchworks Technologies (Costa Rica), Modus Create (Costa Rica), Creativa Consultores (El Salvador), R/Cs (Costa Rica), Applaudo Studios, Rootstack (Panamá), Paralleldevs (Costa Rica), Lionmane Software, Inc. (Guatemala) and Central American Software Services (El Salvador).<sup>13</sup>

These companies have some clients in their own countries that are pushing for an increase in the use of technologies in the population and the growth of e-commerce, however, most of their customers are outside Central America, that rely on them for app development at a lower costs.

Another important difficulty in the region is judicial security; few countries have approved laws to protect and incentive technology investment and business. El Salvador is about to approve new laws in that matter, and it has become so important that it was a key part in the political campaigns. If they do so, the country will match up with Guatemala, Costa Rica, and Panama that have important advances.

The Caribbean has these laws since 1999 with the approval of the Electronic Transactions Act of Barbados, that has been followed for 13 other island states. Another important law was the Electronic Communications and Transactions Act of Bahamas in 2003 that was also created in the Cayman Islands. And finally, the Electronic Commerce and Digital Signature Law has already been created in Dominican Republic and Puerto Rico. The growth of e-commerce and the expansion of computer technologies are pressuring governments to move faster in legislation.<sup>8</sup>

Fiber optic is still important in the region, but in 2019 there was a decrease of \$8 million in comparison to the previous year. Costa Rica is the main importer along with Panamá, but last year they decreased their purchase in approximately 30% while Honduras, Guatemala and El Salvador increased them in more than 60%.

The Caribbean has multimillionaire submarine optic fiber like the East Caribbean Fiber System (ECFS) that interconnects 14 eastern Caribbean islands. They are also looking to develop, with the help of the World Bank, 5G technology in the midterm. China is the most important supplier.<sup>6</sup>

Internet access has increased but it basically still moves through poverty, in 2018 and 2019 the migrant caravans of hundreds of thousands of Central Americans walking to the U.S. were registered through the cellphones of the people that lived in extreme poverty but still managed to have a profile in a social network.

### Social and Gender Inequality

Computer technology moves at quite different speeds in Central America and the Caribbean, and it is not always related to connectivity. The gender inequalities are evident and in countries like Honduras or Haiti men can have over three times more access to technology than women living in rural areas. Even in countries with higher range for technology access like Costa Rica there has been a drop in the participation of women working in computer technology companies from almost a 40% of participation 10 years ago to less than 30% three years ago.<sup>8</sup>


Central American and Caribbean women have important responsibilities over their shoulders, almost three of every five houses are supported by single women that do not have the money and cannot access credit to develop entrepreneurship companies, and is the reason why almost 70% of women work in the service or the clothing sector for very low salaries.

It is ironic but in comparison with other parts of the world, in Latin America women have a higher rate of attainment of master's or doctorate degrees. They graduate but there are fewer spaces for them in the markets. Gender inequality has also made its way to the computer technology world in the region. Several universities have started to develop studies on Cyber Feminism to identify these inequalities and to find ways to face them with economic and political changes that avoid the reproduction of the same historical behavior of other sectors.<sup>1</sup>

Cyber feminism in the region has



**Computer technology moves at quite different speeds in Central America and the Caribbean, and it is not always related to connectivity.**







## For the island-states, digital connectivity has become as important as roads and electricity.



also focus on using computer technologies to reduce the economical gap that affects more women than men. Projects like Women's P2P Network in Haiti has created new markets for thousands of women that can reach new customers, find alliances, and built networks with women all over the country using voice commands to overcome illiteracy. Or KOFIVIV that helped gathered thousands of testimonies of women that had been victims and abused and were afraid to speak out. El Salvador and Nicaragua have created intensive programs to help poor rural women to get close to computer technologies and apply it in their daily activities.

In Guatemala, georeferencing is being used to alert on places with high numbers of attacks against women, and in Cuba is used to help single mothers to look for jobs. In the Dominican Republic, CIPAF is helping international and national companies to identify professional women specialized in computer technology to help them get jobs and are overseeing that salaries of men and women in the same positions are equal.

In the Caribbean, many governments are applying a sort of tax called the Universal Service that charge telecommunication and technology international companies and uses the resource to help people with the lowest incomes to be able to access technology and education. Many of these programs are focused on helping girls to reduce social and gender inequality in technology access.

### Unifying the Islands

In the Caribbean, development is more spread out than in Central America, but it is also a very unequal region. The nations (island-states) of the Caribbean region are scattered over an area of more than 27.5 million square kilometers. These small, developing countries are mostly isolated from each other by the Caribbean Sea and Atlantic Ocean. As such, they have greater Internet connectivity challenges than do mainland developing nations.<sup>2</sup>

The islands depend on submarine cable for Internet connectivity via a combination of fiber-optic, coaxial and copper cables, and fixed or mobile

wireless networks. Second-generation (2G) mobile/cellular networks cover most of the population, providing basic telecommunications (voice and text messaging) services. Increasingly, newer third-generation (3G), fourth-generation (4G) and long-term evolution (LTE) mobile technologies that support mobile broadband Internet are being deployed. Internet access varies, in Haiti 11% of people are connected, and in Barbados and the Bahamas 80% are connected. There are great advances in computer technology in Barbados, Trinidad and Tobago, Bahamas, the Dominican Republic, and Jamaica, which are countries that are increasing their investment in new software companies and presence of worldwide companies.

The World Bank is financing several programs to unify various investments in the different countries: "Internet to unify the islands" they assure. They are focusing on computer science, Internet access, and engineering to develop a stronger market in the Caribbean and attract international investors to tourism, banking, and technology.


The World Bank's Caribbean Regional Communications Infrastructure Program (or CARCIP, for short), is a program that aims to promote digital skills training and business development. The goal is to create jobs while strengthening the countries' entrepreneurial base. With a well-trained population and a solid digital infrastructure, the Caribbean could position itself as a destination for IT services. The project offers resources to train and certify young people in courses of study ranging from software and apps development to database management and Web development.<sup>12</sup>

Approximately 40% of the Caribbean small States population still lacks access to the Internet, and affordability remains a challenge even for those connected. The digital economy is expected to reach 25% of global GDP in less than a decade, making it one of the main sources of growth and job creation. Moreover, studies show a significant positive relationship between digital technology adoption and GDP growth.<sup>7</sup>

Many developing countries have made ambitious strides to take advantage of these opportunities. In



underprepared to apply for employment for housework or to the technology industry and live in countries that are struggling to become attractive to this kind of investment, or to develop computer technologies by themselves. Their possibilities of success vanish by the second.

Millions of young men and women are missing out on the technological development experienced in other regions of the world due to the lack of education and working opportunities. As in the Costa Rica example, education and investment seem to be the keys that could allow Central America and the Caribbean to change their current situation. 

#### References

1. Bonder, G. Las nuevas tecnologías de información y las mujeres: reflexiones necesarias. Unidad Mujer y Desarrollo Proyecto CEPAL-GTZ "Institucionalización del Enfoque de Género en la CEPAL y Ministerios Sectoriales." Santiago de Chile, June 2002; [https://repositorio.cepal.org/bitstream/handle/11362/5894/1/S026404\\_es.pdf](https://repositorio.cepal.org/bitstream/handle/11362/5894/1/S026404_es.pdf)
2. Fonseca-Hoeve, B. et al. Unleashing the Internet in the Caribbean, Removing Barriers to Connectivity and Stimulating Better Access in the Region. Caribglobal Data Services, ICT Pulse Consulting and Internet Society, Feb. 2017; [https://unctad.org/meetings/en/Contribution/dtL\\_eWeek2017c06-isoc\\_en.pdf](https://unctad.org/meetings/en/Contribution/dtL_eWeek2017c06-isoc_en.pdf)
3. Gallego, J.M. and Gutiérrez, L.H. ICTs in Latin America and the Caribbean: Stylized Facts, Programs and Policies Knowledge Sharing Forum on Development Experiences: Comparative Experiences of Korea and Latin America and the Caribbean (2015); <https://www.iadb.org/en>
4. Interview with Arianna Tristán, Director of Innovation and International Affairs of Costa Rica Industrial Chamber. 2020.
5. Interview with Rosemary Hernandez, International Coordinator Costa Rica National University, 2020
6. Loucks, J. et al. Future in the balance? How countries are pursuing an AI advantage. Deloitte Insights. 2019; <https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/ai-investment-by-country.html>
7. Michalczewsky, K. and Ramos, A. E-regulación en América Latina. Interamerican Development Bank (2017); <https://conexionintal.iadb.org/2017/03/08/comercio-electronico-en-america-latina-la-brecha-normativa-2/?lang=en>
8. Morgan, B. List of E-Commerce Laws in the Caribbean. 2020; <https://www.bartlettmorgan.com/2020/04/10/list-of-e-commerce-laws-in-the-caribbean/>
9. Mujeres en la economía digital. Superar el umbral de la desigualdad. XII Conferencia Regional SOBRE LA MUJER de América Latina y el Caribe. 2013; [https://repositorio.cepal.org/bitstream/handle/11362/16561/1/S2013579\\_es.pdf](https://repositorio.cepal.org/bitstream/handle/11362/16561/1/S2013579_es.pdf)
10. OECD. *Shaping the Digital Transformation in Latin America: Strengthening Productivity, Improving Lives*. OECD Publishing, Paris, 2019; <https://doi.org/10.1787/8bb3c9f1-en>.
11. Rodríguez L., Germán, L., Carnota, R. Historias de las TIC en América Latina y el Caribe: Inicios, desarrollos y Ruptura, Fundación Telefónica. 2015.
12. Sayed, T. and Habalian, R. Digitally transforming the Eastern Caribbean, 2019 World Bank; <https://blogs.worldbank.org/latinamerica/digitally-transforming-eastern-caribbean>
13. *The Manifest* (2020); <https://themanifest.com/app-development/companies/latin-america>

**Gerardo Torres Zelaya** is a journalist specializing in political and economic analysis. He is based in Honduras.

© 2020 ACM 0001-0782/20/11

fact, many low to lower middle-income countries have embraced digital technologies to transform public services and reduce transaction costs for their citizens. In the last years all the island-states in the Caribbean have been focusing on the creation of State programs and new laws to help the development of information and communication technologies (ICTs). Large and medium-sized enterprises generally have access to the Internet, but the adoption of advanced ICTs is low for all firms in these economies, and small and micro enterprises lag way behind. The backwardness in ICT adoption is exacerbated when only a small fraction of society has high connectivity broadband. Only two of the 24 jurisdictions surveyed appear to be without laws targeted at e-commerce: those are Cuba and Guyana (as of April 2020).<sup>3</sup>

For Grenada, St. Lucia, St. Vincent and the Grenadines, broadband Internet access is a way to tie small, somewhat isolated island populations to each other and to the rest of the world. Fast, on-demand digital connections to business opportunities, information, even to friends and relatives is even more crucial when you are perched on a small island in the Caribbean Sea. For the island-states, digital connectivity has become as important as roads or electricity.

And while many Caribbean countries are now middle-income countries, unemployment rates have stayed high, above 20% in St. Vincent and the Grenadines and St. Lucia. Better

broadband access should help the poorest to share the prosperity of their neighbors. This regional partnership was kicked off with a workshop co-hosted by the Eastern Caribbean Central Bank (ECCB) and the World Bank. This was the first time ever such a comprehensive and holistic dialogue on the topic took place between the World Bank and regional/country partners.<sup>2</sup>

This engagement will build on the ongoing Caribbean Regional Communications Infrastructure Program, supporting the deployment of over 1,200km of terrestrial and submarine fiber-optic cables to strengthen digital connectivity in select Eastern Caribbean countries. It also complements the Grenada Digital Governance for Resilience project, approved by the World Bank Board in August 2019. The first Caribbean project of this nature seeks to enhance the efficiency, usage, and resilience of select government services in Grenada.

#### The Keys to Succeed

In 2020, the COVID-19 pandemic and the extended quarantine in the Central American and the Caribbean countries has brought a burst in the growth of e-commerce especially in the middle and the upper class, but it has meant a whole new series of problems for the poor (50% of the population) whom—for example—in most of these countries, have not had access for public education during the pandemic, and probably will not have for the remainder of the year. They are



BY CLAUDIO DELRIEUX, VIRGINIA BALLARÍN,  
CRISTIAN GARCÍA BAUZA, AND MARIO A. LÓPEZ

# Imaging Sciences R&D Laboratories in Argentina

WE USE THE term *imaging sciences* to refer to the overarching spectrum of scientific and technological contexts which involve images in digital format including, among others, image and video processing, scientific visualization, computer graphics, animations in games and simulators, remote sensing imagery, and also the wide set of associated application areas that have become ubiquitous during the last decade in science, art, human-computer interaction, entertainment, social networks, and many others. As an area that combines mathematics, engineering, and computer science, this discipline arose in a few universities in Argentina mostly in the form of elective classes and small research projects in electrical engineering or computer science departments. Only in the mid-2000s did some initiatives aiming to

generate joint activities and to provide identity and visibility to the discipline start to appear. In this short paper, we present a brief history of the three laboratories with the most relevant research and development (R&D) activities in the discipline in Argentina, namely the Imaging Sciences Laboratory of the Universidad Nacional del Sur, the PLADEMA Institute at the Universidad Nacional del Centro de la Provincia de Buenos Aires, and the Image Processing Laboratory at the Universidad Nacional de Mar del Plata.

The Imaging Sciences Laboratory<sup>a</sup> of the Electrical and Computer Engineering Department of the Universidad Nacional del Sur Bahía Blanca began its activities in the 1990s as a pioneer in Argentina and Latin America in research and teaching in computer graphics, and in visualization. The facility currently is staffed by six National Research and Technology Council (CONICET) fellows (Claudio Delrieux, Alejandro Vitale, Felix Thomsen, Natalia Revollo, Marina Cipolletti, and Noelia Revollo), plus three postdocs and 13 Ph.D. candidates, who are actively researching novel image analysis methods including multifractality, complexity theory, and deep learning in the hope of generating breakthroughs in research contexts including three-dimensional (3D) medical images,<sup>1</sup> 3D shape analysis,<sup>2</sup> biometrics,<sup>3</sup> biomedical signal analysis,<sup>4</sup> microscopy,<sup>5</sup> remote sensing,<sup>6</sup> satellite imagery,<sup>7</sup> and environmental monitoring,<sup>8</sup> while maintaining research and development activities in scientific visualization and computer graphics.

In Figure 1, we show novel 3D texture analysis techniques applied to brain MRIs in mild and moderate Alzheimer's disease patients (and in comparison with a control population of subjects of similar age and condition). The proposed operators detect compromised areas of the subjects' brains that correlate to their actual

<sup>a</sup> [www.imaglabs.org](http://www.imaglabs.org)

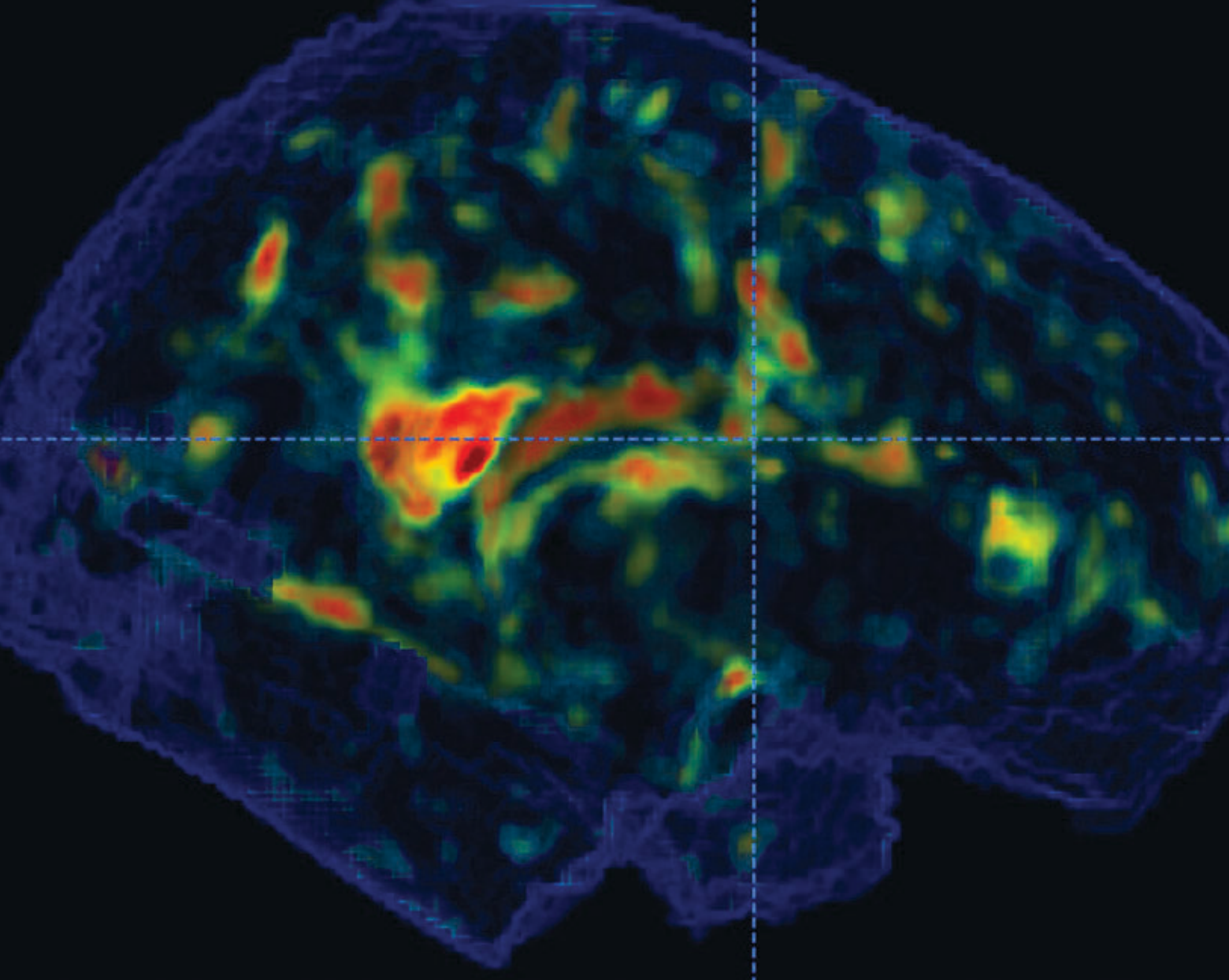


IMAGE COURTESY OF THE IMAGING SCIENCES LABORATORY/IMAGI.ABS.ORG

neurocognitive impairments, and are able to identify compromised areas even before the subjects actually experience any abnormality with high statistical significance.<sup>1</sup> In Figure 2, we show the results of a demographic analysis of body shape and obesity, using inexpensive smartphones to acquire 3D body shapes. Specialists scanned a total of 250 people using LIDAR devices and measured them using standard anthropometric procedures. Also, we developed a deep-learning-based photogrammetric procedure to compute point clouds from short-range videos taken with smartphones; the anthropometric measurements were similar in quality and accuracy to those obtained with LIDAR. An obesity and body-shape analysis of this population shows that Body Mass Index (BMI) and similar indices may be misleading, and our

study proposes a different assessment strategy.<sup>3</sup>

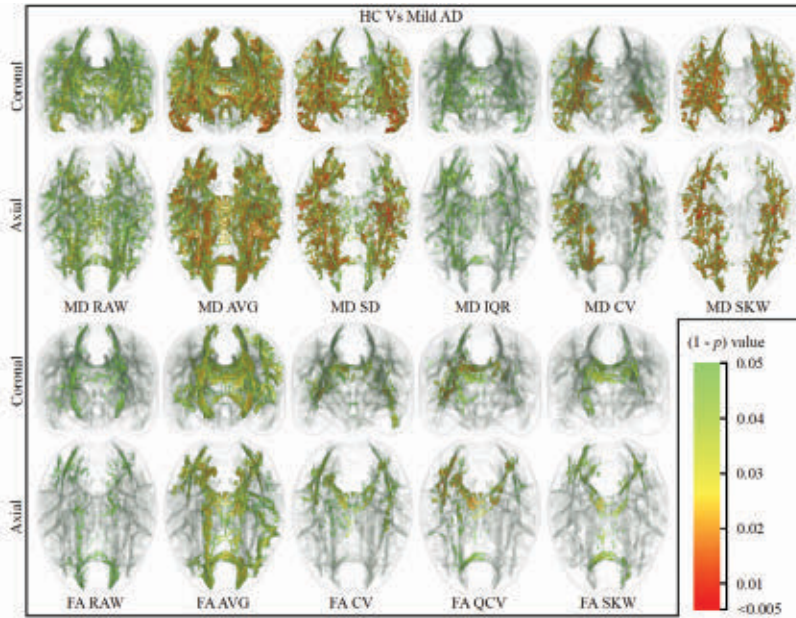
Apart from these research activities, the Imaging Sciences Laboratory is a nationwide reference in I+D+i (*Investigación, desarrollo, e innovación*, or research, development, and innovation) and technological transfer to other stakeholders. Since 2006, the lab has participated significantly in the development and deployment of innovative products for more than 20 companies and institutions, and has been awarded several national innovation prizes. Among this group of applied projects, we note the development of Agroinfintiy, an information management platform commissioned by the Chamber of Crop Producers of Bahía Blanca,<sup>b</sup> which includes the monitoring and productive assess-

<sup>b</sup> [www.bcp.org.ar](http://www.bcp.org.ar)

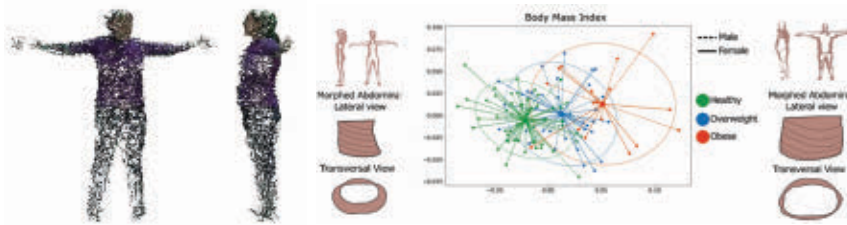
ment of more than 200.000 sq.km. of agro-productive areas. This platform is able to register and process relevant indicators from global scale to producer-focused scales, using satellite and airborne imagery, weather information, and field observations, taking into account all the economic variables to elaborate a full recommendation. The group is involved in several ongoing collaborations with internationally renowned groups, including the Zentrum für Virtual Reality und Visualisierung (VrVis) and Vienna University of Technology (TU Wien) in Austria, the Universidade Federal Rural de Pernambuco and Universidade Federal de Alagoas in Brasil, the Universidad de Valladolid in Spain, the University of Denver in the U.S., and several other national centers in Argentina. The group also regularly organizes graduate classes



**Figure 1. Per-voxel statistical significance of different novel three-dimensional texture analyses and their correlation to different degrees of Alzheimer’s disease. The operators detect compromised areas in the subjects’ brains that correlate to their actual neurocognitive impairments.<sup>1</sup>**



**Figure 2 (Left): A point cloud generated with handheld devices using photogrammetry and deep learning. (Right): A 250-person body shape and overweight analysis shows that BMI and similar indices may be misleading, and proposes different assessment strategies.<sup>3</sup>**



**Figure 3. Subway train simulators for the Buenos Aires city network.**



with invited professors from these and other internationally recognized research groups.

**The PLADEMA Institute**

The PLADEMA Institute<sup>c</sup> depends on the School of Sciences of the Universidad Nacional del Centro de la Provincia de Buenos Aires (UNCPBA), the National Atomic Energy Commission (CNEA) and the Scientific Research Commission (CIC) of the Buenos Aires Province. It began its activities in 1997, standing out for providing solutions to applied problems using a multidisciplinary approach, providing its main contributions in the areas of simulation, mathematical modeling, high-performance computing, and graphics computing. These fields apply cross-cutting methodological tools to tackle different problems using complementary approaches, fostering the development of novel solutions of great impact from the social and productive point of view at the local, regional, and national levels. Currently, PLADEMA has six research teams focused on specific topics. The whole team consists of more than 100 members (20 Ph.D.’s, 18 doctoral fellows, four support personnel, and more than 60 interns). Three groups are actively working in image processing or computer graphics: Yatiris, researching problems related to medicine, including the processing of medical images; MediaLab, which specializes in computer graphics, virtual reality, and the development of training simulators; and RedDot, which focuses on video analysis, computer vision, and signal processing. The main researchers of these groups are Marcelo Vénere, Alejandro Clause, Ignacio Larra-bide, Cristian García-Bauza, and Juan D’Amato.

Some results from the Yatiris team in image processing include: automatic segmentation in magnetic resonance imaging (MRI) and computed tomography (CT) images for the detection and discretization of arteries and intravascular ultrasound images in order to perform hemodynamic modeling, and the use of deep learning techniques to detect retino-

<sup>c</sup> www.pladema.net



blastomas in ocular images.<sup>16</sup> Interesting advances were achieved in training areas with the implementation of an abdominal ultrasound simulator<sup>17</sup> and the construction of anatomical tables for the teaching of anatomy.

The MediaLab team is a mainstay within Argentina when it comes to work in computer graphics and virtual reality. Innovative techniques to achieve highly realistic modeling effects have been developed by the group for use in real-time applications and video games;<sup>18</sup> also, a computing platform that allows the real-time management of multiple projection surfaces was built, based on which the first Cave Automatic Virtual Environments (CAVEs) were developed in Argentina; MediaLab's multidisciplinary team achieved the successful CAVE implementation using Virtual Reality techniques. Interactive applications using video game concepts applied to the social and cognitive development of children with intellectual disabilities also have been used at the lab to allow learning through play.<sup>19</sup>

Finally, the RedDot team specializes in real-time image processing applied to computer vision and security applications. Through the use of deep learning techniques, the team has developed algorithms able to monitor scenes and detect the behavior of certain objects.<sup>20</sup> Solutions have been implemented for urban traffic, movement of people, and detection of equipment failures in industrial plants.

PLADEMA is a strong actor in the transfer and development of applied technological products. Since 2001, PLADEMA has implemented more than 30 innovative products for Ministries, other government entities, and private companies, winning many national innovation awards. A good example of this kind of applied project is the development of a simulator covering the entire subway network of the city of Buenos Aires to train the drivers of subway trains how to handle system failures (Figure 3). Training with the simulator enables drivers to practice solving various types of faults on a varied set of trains, within a simulated virtual environment that emulates real processes. The project faced great technological challenges at

the national and international levels, including the accurate modeling of the physical behavior of the trains, simulation of the movement of large crowds, representation of many kilometers of tunnels and tracks, and of the geometry of more than 100 railway stations. Furthermore, everything must work in real time.

The implementation of the RUBIKA platform is another success story that includes the construction of nine CAVEs, installed at different institutions in our country, along with the development of software packages for training in several areas, such as teaching in petroleum engineering careers, and the design of nuclear power plants (Figure 4). The goal was to design, build, and fully develop a local solution for several applications with the same quality as its international equivalents, taking advantage of working with local knowledge and lower costs. This required solving complex technical challenges concerning the hardware and software that make up the environment using low-cost projectors so the device can be easily maintained, and required the development of a software platform with substantial added value in terms of calibration, distributed processing, and optical corrections. The structure has to allow the adjustment of the four surfaces (three walls and the floor) and the calibration of the projectors

to ensure that the projected images are less than one pixel in error. Each surface is handled by a separate computer, and all four processes must run synchronously. Finally, the position of the user's head must be detected in real time to generate the correctly corresponding projection.

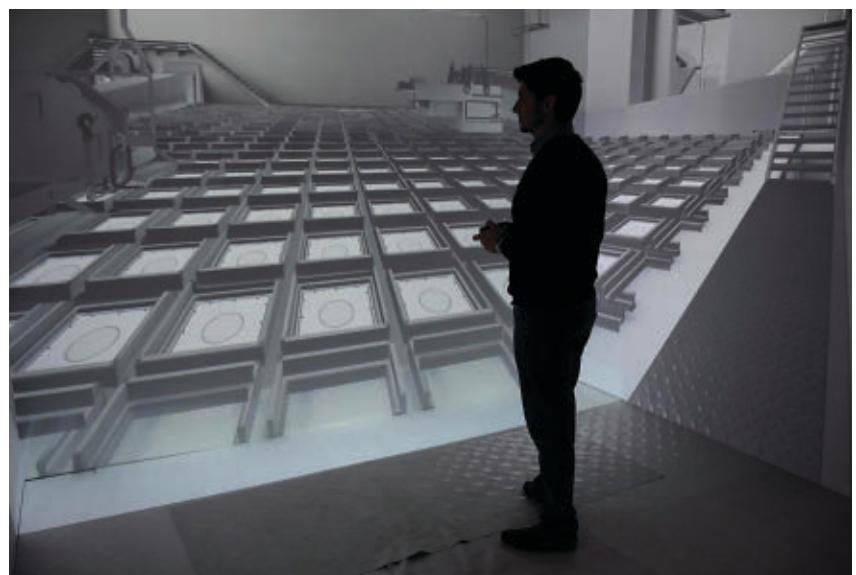
PLADEMA keeps close collaborative relationships with internationally well-known groups and with several R&D institutions in Argentina. Also, the institute regularly holds postgraduate classes given by professors from these and other outstanding research groups. PLADEMA is also a benchmark for local and regional training offers, especially regarding the implementation of academic proposals such as technical and short careers, which are planned and designed together with government and business stakeholders, such as the Municipality of Tandil City and the Chamber of Companies of the Computer Pole of Tandil (CEPIT).

### The Image Processing Laboratory

The Image Processing Laboratory (PILab) is part of the Institute of Scientific and Technological Research in Electronics (ICYTE)<sup>d</sup> of the Universidad Nacional de Mar del Plata (UNMDP) and CONICET. It started in 1987 as part of the Image Analysis and

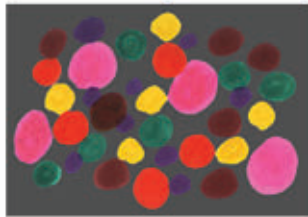
<sup>d</sup> <https://icyte.conicet.gov.ar>

**Figure 4. RUBIKA is a software and hardware platform that generates immersive virtual environments for training.**

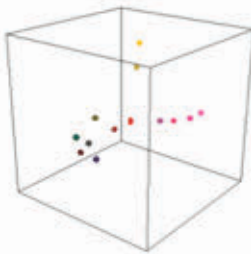




**Figure 5. W-operators of the mathematical color morphology using diffuse definitions of color spaces: example of erosion and dilation applied to a color image. (a) Original color image. (b) Situation of the representative crisp colors in the RGB color space. (c) Volumes of colors in the 0.5-cut for the fuzzy colors yellow, blue, green, and gray obtained from the Voronoi diagram in the RGB cube. (d) Dilation. (e) Erosion.**



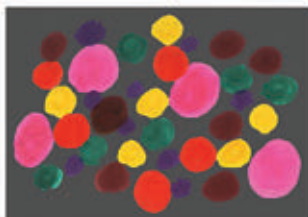
(a)



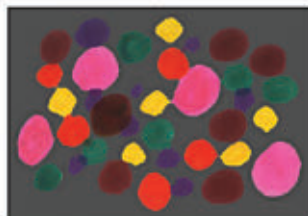
(b)



(c)



(d)



(e)

Coding Group (COANIM) of UNMDP, which later was dissolved since most members moved to the private sector. The remaining members created the Image Processing Group in the early 1990s. This group was strengthened with the addition of CONICET and CIC fellows, later formally becoming the PILab.

Since its creation, PILab has been nourished by numerous scholarships of all kinds, undergraduate students, CONICET and CIC Doctoral Scholarships, always maintaining a high number of junior participants (currently 50%). More than half of the scholarship holders come from abroad, mainly from Ecuador and Colombia. Argentine scholarship holders typically decide to join as CONICET fellows, thus staying at the lab. The PILab currently has 20 members: eight seniors (Virginia Ballarin, Juan Pastore, Guillermo Abras, Eduardo Blotta, Agustina Bouchet, Marcel Brun, Diego Comas, and Inti Pagnucco), two post-docs, six Ph.D. candidates, and four interns. Its initial area of expertise was mathematical morphology, pioneering this topic in Argentina. Over time, the lab's goals shifted to the development of applications in medical imaging. Currently, three main research approaches coexist at the lab:

**1. Design of color morphological operators and fuzzy morphological operators.** Color is a very important visual feature in computer vision and image processing. The extension of grayscale image algorithms to color

is not always direct.<sup>9,10</sup> The PILab is developing theoretical advances in this direction, defining color morphological operators in a novel fuzzy color space (Figure 5).<sup>11,12</sup> The automatic design of w-operators in gray levels and in color space is not only a theoretical approach of the lab, but these techniques also are being used to develop medical applications.<sup>13</sup>

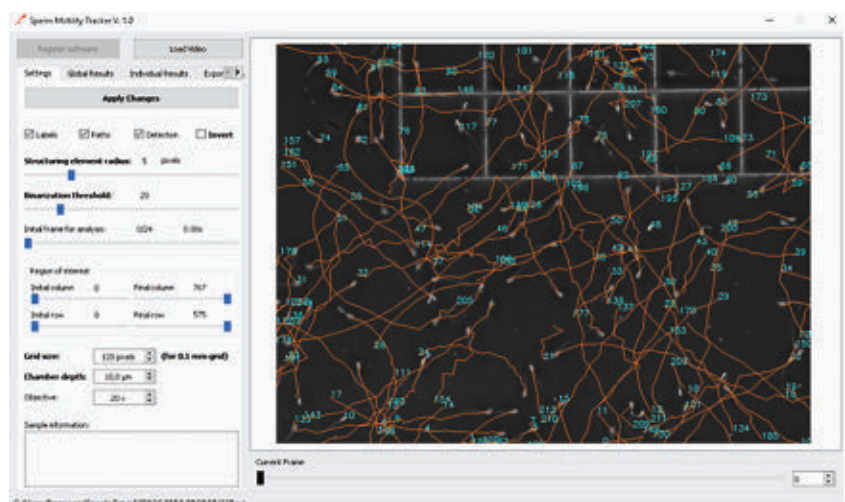
**2. Deep learning models for segmentation and knowledge discovery in medical images.**

From previous experience in discovering knowledge from data clustering using fuzzy systems,<sup>14</sup> the new challenge regarding this paradigm is to be able to answer important questions about networks based on deep learning. Many of these questions have not been analyzed in the research literature. This research will allow us to study in depth the type of knowledge that can be extracted from answering questions such as:

- ▶ Is it possible to use the data representation contained in the internal parameters of the networks to extract interpretable knowledge?
- ▶ What is the meaning of these parameters, and how can we take advantage of them to extract knowledge?
- ▶ What does the internal data of the network describe regarding the problem addressed?
- ▶ How are changes in the input data reflected in the internal parameters of the networks?

**3. Tracking moving objects in image sequences.** Object tracking systems in image sequences have been

**Figure 6. Sperm motility tracker software.**





restricted to rigid objects. The image sequences from biological experimentation do not fit these models, which is why the PILab has increased its research efforts in tracking objects with other modes of movement, developing efficient methods to segment and quantify these objects.<sup>15</sup>

Apart from these research activities, the PILab has participated in various technological transfer initiatives ranging from fingerprints restoration for the Argentine Forensic Anthropology Team (EAAF), and DNA Images for Human Identification software for the Mar del Plata Human Genetics Foundation, to Sperm Motility Tracker software (Figure 6) for the Biological Research Institute (IIB).


The group maintains close collaboration with internationally renowned groups, including the College of Engineering of Texas A&M University, U.S.; the Image processing Lab of the Instituto de Matemática e Estatística of the Universidade de São Paulo, Brazil; the CEATIC Center for Advanced Studies in Communication Technologies of the Universidad de Jaen, Spain, and the Uncertainty and Imprecision Modelling in Decision Making group of the Universidad de Oviedo, Spain. The PILab also maintains close collaboration with local centers of excellence to which the PILab transfers technology, such as the Institute of Biological Research (IIB-CONICET), the Plant Physiology Laboratory of the INTA Balcarce, and especially with the Institute of Research in Technology and Materials Science (INTEMA).

### Joint Activities and Conclusion

Aiming to gain visibility as a discipline, these three groups jointly sponsor other activities which include the organization of a yearly School and Workshop in Imaging Sciences (ECIMAG),<sup>e</sup> and a joint Ph.D. program in medical imaging. The ECIMAG was held from 2008 until 2014, after which the financial situation of the scientific communities in Argentina suffered a huge shortage. During these years, the event was organized by universities in some of the major cities of Argentina, including Buenos Aires, Tandil, Bahía

Blanca, and Santa Fe. Apart from regular paper and poster presentations, the school would bring between two and four renowned invited professors from international research centers to deliver intensive courses focused on advanced topics related to image processing.

During these events, it became apparent that medical imaging was an important common research topic of the three research groups, while all three faced similar difficulties in developing Ph.D. programs in the area, due to the lack of critical mass in faculty. For this reason, a project for developing a joint Ph.D. program in the topic was presented and finally approved. In late 2018, the degree was opened at the Universidad Nacional del Sur, obtaining official accreditation from the corresponding Ministry Agency (CONEAU). Currently there are three candidates in the program at that university, with other candidates submitting applications. During 2019, the degree was opened at the Universidad Nacional de Mar del Plata and the Universidad Nacional del Centro de la Provincia de Buenos Aires, and their respective accreditations are due in the months to come.

While the field of imaging sciences started as a fringe area in engineering departments, it has matured and developed into a discipline in its own right in Argentina, as evidenced by the accomplishments of the three research groups. In spite of financial limitations, the groups have grown in research, development, and extended collaborations internationally, making the future of imaging science in Argentina look bright indeed. 

### References

1. Thomsen, F.S.L., Delrieux, C.A., and de Luis-García, R. Local texture descriptors for the assessment of differences in diffusion magnetic resonance imaging of the brain. *Intern. J. Computer-Assisted Radiology and Surgery* 12, 3 (2018), 389–398.
2. Navarro, P., et al. Body shape: Implications in the study of obesity and related traits. *American J. Human Biology* 32, 2 (2020).
3. Cintas, C., Quinto-Sánchez, M., Acuña, V., Paschetta, C., de Azevedo, S., and Delrieux, C. Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks. *IET Biometrics* 6, 3 (2017), 211–223.
4. Avila, F., Delrieux, C., and Gasaneo, G. Complexity analysis of eye-tracking trajectories. *The European Physical J.* 92, 12 (2019).
5. Revollo, N.V., Delrieux, C.A., and González-José, R. Set of bilateral and radial symmetry shape descriptor based on contour information. *IET Computer Vision* 11, 3 (2018), 226–236.
6. Sarmiento, G.N.R., Cipolletti, M.P., Perillo, M.M.,

- Delrieux, C.A., and Perillo, G.M.E. Methodology for classification of geographical features with remote sensing images: Application to tidal flats. *Geomorphology* 257 (2018), 10–22.
7. Cipolletti, M.P., Genchi, S.A., Delrieux, C., and Perillo G.M.E. An approach for estimating border length in marine coasts from MODIS data. *IEEE Geoscience and Remote Sensing Letters* 17, 1 (2020), 8–12.
8. Genchi, S.A., Vitale, A.J., Perillo, G.M.E., Seitz, C., and Delrieux, C.A. Mapping topobathymetry in a shallow tidal environment using low-cost technology. *Remote Sensing* 12, 9 (2020).
9. Bouchet, A., Alonso, P., Pastore, J.I., Montes, S., and Díaz, I. Fuzzy Mathematical Morphology for color images defined by fuzzy preference relations. *Pattern Recognition*. Elsevier (Dec. 2016), 720–733, DOI: 10.1016/j.patcog.2016.06.014.
10. Pastore, J., Bouchet, J.A., Brun, M., and Ballarin, V. New windows-based color morphological operators for biomedical image processing. *J. Physics*. 705, 1 (Apr. 2016).
11. Bouchet, A., Colabella, L., Omar, S., Ballarre, J., and Pastore, J. Processing of microCT implant-bone systems images using fuzzy mathematical morphology. *J. Physics* 705, (Apr. 2016).
12. Bouchet, A., Pastore, J.I., Marcel Brun, M., and Ballarin, V.L. Compensatory fuzzy mathematical morphology. *Signal, Image and Video Processing* 11, 6 (Sept. 2017), 1065–1072. Springer.
13. Pastore, J.I., Brun, M., Bouchet, A., and Ballarin, V.L. Color morphological reconstruction as a tool for microscope cell images. *IJFMBE Proceedings* 60 (2017), 312–315. Springer International Publishing.
14. Comas, D.S., Pastore, J.I., Bouchet, A., Ballarin, V.L., and Meschino, G.J. Interpretable interval type-2 fuzzy predicates for data clustering: A new automatic generation method based on self-organizing maps. *Knowledge-Based Systems* 133, 1 (Oct. 2017), 234–254. Elsevier.
15. Imbachí, F.B. et al. Objective evaluation of ram and buck sperm motility by using novel sperm tracker software. *Reproduction J.* (May 22, 2018); DOI: 10.1530/REP-17-0755.
16. Orlando, I., Van Keer, K., Breda, J.B., Manterola, H., Blaschko M., and Clausse, A. Proliferative diabetic retinopathy characterization based on fractal features: Evaluation on a publicly available data set. *Amer Assoc Physicists Medicine Amerinst Physics* 44, 12 (2017), 6425–6434.
17. Vitale, S., Orlando, I., Iarussi, E., and Larrabide, I. Improving realism in patient-specific abdominal ultrasound simulation using CycleGANs. *Intern. J. Computer-Assisted Radiology and Surgery*, 2019, Springer.
18. D'Amato, J.P., García Bauza, C., Lazo, M., and V. Cifuentes, V. Tridimensional Scenes Management and Optimization for Virtual Reality simulators. *Advances in Intelligent Systems and Computing* 444 (2016), 243–252. Springer-Verlag, Heidelberg.
19. Contreras, M., García Bauza, C., and Santos, G. Videogame-based tool for learning in the motor, cognitive and socio-emotional domains for children with intellectual disability. *Entertainment Computing* 30 (2019), DOI: 10.1016/j.entcom.2019.100301.
20. Domínguez, L., D'Amato, J.P., Pérez, A., Rubiales, A., and Barbuzza, R. A GPU-accelerated LPR algorithm on broad vision surveillance camera. *J. Information Systems Engineering & Management* 3 (2018), 1–7.

**Claudio Delrieux** is a professor in the Departamento de Ing. Eléctrica y Computadoras, Universidad Nacional del Sur, Argentina, and Research Fellow in Consejo Nacional de Investigaciones Científicas y Tecnológicas/CONICET.

**Virginia Ballarín** is a professor with the ICYTE CONICET–CIC and on the Facultad de Ingeniería, Universidad Nacional de Mar del Plata, Argentina.

**Cristian García Bauza** is a professor with PLADEMA CNEA – CICCPBA – Facultad de Cs. Exactas, Universidad Nacional del Centro de la Prov. de Buenos Aires, Argentina, and Research Fellow in Consejo Nacional de Investigaciones Científicas y Tecnológicas/CONICET.

**Mario A. López** is Distinguished Professor at the Ritchie School of Engineering and Computer Science of the University of Denver, CO, USA.

<sup>e</sup> [www.ecimag.org](http://www.ecimag.org)



BY ELIAS P. DUARTE JR., RAIMUNDO J. A. MACÊDO,  
ELIANE MARTINS, AND SERGIO RAJSBAUM

# A Tour of Dependable Computing Research in Latin America

COMPUTING TECHNOLOGY HAS become pervasive and with it the expectation for its ready availability when needed, thus basically at all times. Dependability is the set of techniques to build, configure, operate, and manage computer systems to ensure that they are reliable, available, safe, and secure.<sup>1</sup> But alas, faults seem to be inherent to computer systems. Components can simply crash or produce incorrect output due to hardware or software bugs or can be invaded by impostors that orchestrate their behavior. Fault tolerance is the ability to enable a system as a whole to continue operating correctly and with acceptable performance, even if some of its components are faulty.<sup>3</sup>

Fault tolerance is not new; von Neumann himself designed techniques for computers to survive faults.<sup>4</sup>



The premiere conference in the area, the IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), held its 50<sup>th</sup> edition in 2020. In Latin America, the first edition of the Brazilian Symposium on Fault-Tolerant Computers (SCTF) was held in 1985 and took place for 18 consecutive years. In 2003 it evolved into the Latin American Symposium on Dependable Computing, which has since been held in multiple countries. Today, research groups are firmly es-



IMAGE BY MAQUETTE PRO

established, and premier international events have been held in the region, such as DSN in Rio de Janeiro in 2015, the ACM Principles of Distributed Computing (PODC) conference in Mexico in 1998, and the IEEE International Symposium on Reliable Distributed Systems, which has been held twice in Brazil, in Florianópolis in 2004, and in Salvador in 2018, among others.


The first research efforts on dependable computing in Latin America

were focused on aerospace systems. Safety is a dependability property that ensures that a system is capable of avoiding catastrophic consequences. It must be assessed for applications that perform life- or material-critical tasks and is thus a strict requirement of aerospace systems. Established in São Paulo, Brazil, in the early 1960s, the National Institute for Space Research (INPE) has since fostered research in multiple fields including space, climate, and computer


science. At the same time, the aerospace industry took large strides in Brazil, which included the creation of airplane maker Embraer. It is no coincidence that the first SCTF was held at INPE in 1985, chaired by Alderico Rodrigues de Paula Jr., who led the project that developed the fault-tolerant computer launched with the SCD-1 satellite in 1993.<sup>a</sup>

<sup>a</sup> [http://www.inpe.br/scd1/site\\_scd/historico.htm](http://www.inpe.br/scd1/site_scd/historico.htm) (Portuguese)





## Replicating servers over computer networks is one of the building blocks commonly applied to improve the reliability and availability of distributed services.



Since then, research in dependable computing in Latin America (LATAM) has covered many topics, including hardware, software, and communications, from both a practical and theoretical perspective. This article gives a general view of LATAM research in the field and highlights some main results. Please note that we have set up a Web page with an accompanying bibliography.<sup>b</sup>

### **An Overview of Dependable Computing Research in LATAM**

The development of dependable systems requires a combination of fault prevention and fault tolerance techniques to achieve the desired level of reliability, availability, and safety, among other attributes. As such, a major focus of research in LATAM has been the design of hardware and software fault-tolerant architectures to protect systems against faults, either accidental or malicious, complemented by dependability assessment analysis. Code inspection, model-checking, and testing have been widely applied on various LATAM projects, and model-based testing has sometimes been used to generate test cases. Special attention has been devoted to fault injection techniques, not only to dynamically verify the efficacy of fault-tolerant mechanisms, but also to estimate parameters used in analytical models for quantitative evaluation.

On the other hand, architecting dependable systems usually requires the understanding and construction of basic common building blocks or patterns to be reused across multiple application scenarios. Replicating servers over computer networks is one of those building blocks commonly applied to improve the reliability and availability of distributed services—the so-called state machine approach. At the core of this approach are basic problems such as consensus, group communication, group membership, and failure detection. The actual characteristics and behavior of the underlying communication networks, computers, system software, applications, and users impose a myriad of distinct challenges for solving these problems, which have also been exten-

sively addressed in research in LATAM, ranging from theoretical foundations to system engineering and tools.

In the computability and complexity theory realm, handling faults, either by masking or recovering from them, brings additional challenges in algorithm design, especially when systems are distributed, not only in understanding computability limits such as the impossibility of solving certain fault-tolerant problems, but also to proposing, analyzing, and proving solutions for problems such as distributed consensus, renaming, and mutual-exclusion, to cite a few. Thus, a great deal of research in LATAM has been dedicated to the search for appropriate system models, data, and control structures to handle these basic fault-tolerant problems.

At the system and networking management level, other challenges arise. Today systems are widely distributed, concurrent, mobile, and often involve the composition of heterogeneous components, usually requiring autonomous coordination and monitoring. Several pieces of research in LATAM addressed these challenges, proposing new approaches to handle typical problems such as distributed diagnosis, message broadcast, and failure detection, among other problems. Furthermore, theoretical as well as practical approaches to modeling, reasoning, and implementing adaptive and self-adaptive dependable systems in such complex distributed scenarios have been proposed, aiming for autonomous properties such as self-optimization, self-organization, and self-management.

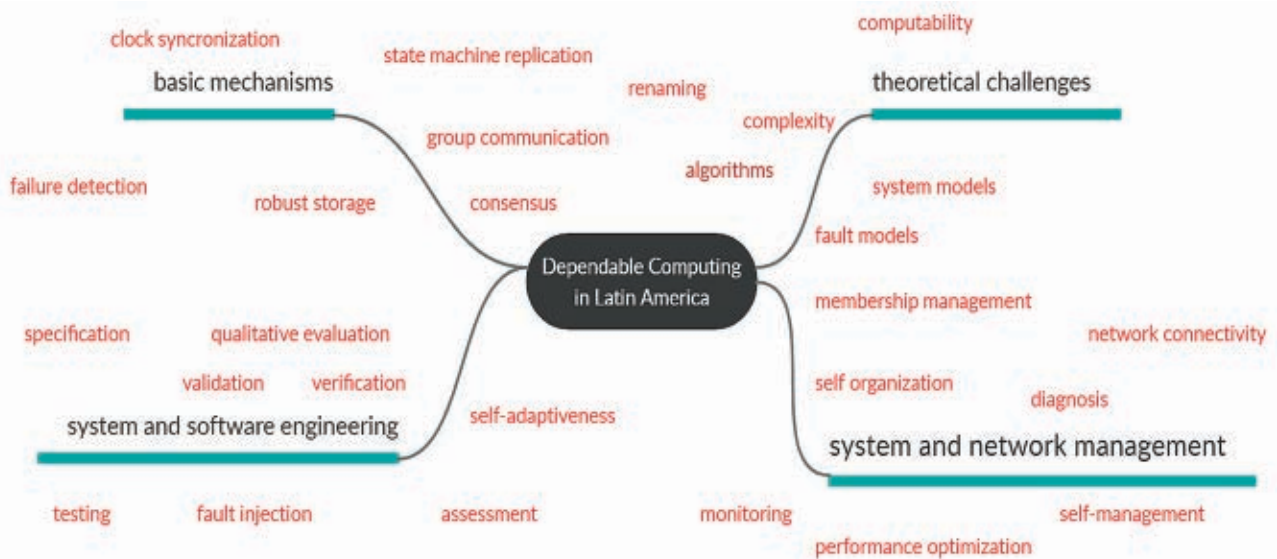
The overall challenges and main topics addressed in dependable computing research in LATAM are shown in the accompanying figure.

### **A Non-Exhaustive Look into Results**

Some of the early dependability projects developed at INPE included robustness testing of satellites, including altitude and orbit control, for which data mining was applied to detect anomalies. Space software was also a major subject, such as the field investigation of errors on space software requirements, done in cooperation with Emilia Villani from the nearby Institute of Aeronautical Tech-

<sup>b</sup> <http://www.inf.ufpr.br/elias/BibDepLA.html>

## Topics of dependable computing research in Latin America.



nology. INPE also fostered projects on integrating model checking and model-based testing for industrial software development. The Brazilian pioneer in this field is Eliane Martins (UNICAMP, Brazil), who has led several projects on dependable computing in cooperation with INPE, in particular with Ana Maria Ambrosio and Fátima Mattiolo-Francisco. One of their projects on a Test Environment with Fault Injection by Software (ATIFS)<sup>c</sup> proposed the combination of conformance testing with fault injection. They also worked on the automation of specification-based test case generation for communication systems, and effective testing strategies for the interoperability and robustness of real-time embedded software. Combining model-driven engineering and model-based testing has been recently explored to cope with dynamically evolving systems, with the contribution of Leonardo Montecchi.

Earlier work on fault injection for dependability validation was developed by Martins with Jean Arlat (France). That work included the development of estimators of the coverage of fault-tolerant mechanisms computed using statistical analysis of data obtained with fault injection. Later, Martins also worked on fault injection strategies based

on reflective programming and on patterns. With Regina Moraes (UNICAMP), Henrique Madeira (Portugal), and Marco Vieira (Portugal), they proposed a strategy based on fault injection for risk assessment. A Java framework to specify fault-loads for fault injection campaigns was proposed by Taisy Weber and Sergio Cecchin (UFRGS, Brazil). They also developed the FIONA tool, a fault injector for dependability evaluation of Java-based network applications.

The development of dependable software based on exception handling mechanisms was pioneered by Cecilia Rubira (UNICAMP), who has worked on different aspects of this problem. A comparison of exception handling techniques for object-oriented software with Alessandro Garcia (PUC-Rio, Brazil), Alexander B. Romanovsky, and Jie Xu (U.K.) is highly cited. An architectural approach for effectively representing and analyzing fault-tolerant software systems has been proposed with Rogerio de Lemos (U.K.) that relies on exception handling to tolerate diverse types of faults. Rubira has worked with Andrea Bondavalli (Italy) on the dependability of dynamic software product lines (SPLs) and SPLs for supporting fault-tolerant composite services.

An evaluation of air traffic controller workloads considering manned and unmanned aircraft systems is

one of several projects on the safety of critical cyber physical systems by João Batista Camargo (USP, Brazil). Other projects include anomaly detection in railway and subway systems. As confidence in safety analysis is essential, Camargo has proposed practical analytical approaches to increase confidence in systems based on programmable logic devices, and on safety-critical software.

The TANGRAM-II tool for modeling system performance and availability, which has been widely used in both academia and industry worldwide, was developed by Edmundo de Souza e Silva (UFRJ, Brazil). Among his many contributions are strategies to compute availability and performability measures of repairable computer systems using randomization.

Armando Castañeda and Sergio Rajsbaum (UNAM, Mexico) are well-known for their work on the theoretical foundations of fault-tolerant distributed systems. Together with Achour Mostefaoui and Michel Raynal (France) they investigated the conditions that identify sets of input vectors for which it is possible to solve consensus despite the occurrence of up to  $t$  process crashes, extended later with Roy Friedman (Israel) to the relationship of asynchronous agreement with error-correcting code. In collaboration with Idit Keidar (Israel), Rajsbaum investigated the cost of solv-

<sup>c</sup> <http://www3.inpe.br/atifs/> (Portuguese)



**The development of dependable systems requires a combination of fault prevention and fault tolerance techniques to achieve the desired level of reliability, availability, and safety.**

ing consensus in failure-free executions. A recent innovative result is on generalizing linearizability to interval-linearizability, allowing specifications of concurrent problems. Robotics has also been a topic of recent research, including an investigation of robots under the Asynchronous Luminous Robots model operating in look-compute-move rounds in connection with shared-memory wait-free algorithms. Together with Pierre Fraigniaud (France) and Corentin Travers (France) they initiated the study of runtime distributed monitors, which while monitoring the correctness of the underlying system, tolerated failures of the monitors themselves.

Castañeda in his Ph.D. work studied the renaming problem, in which processes start with unique input names from a large space and must choose unique output names taken from a smaller name space. This corrected a result of a paper that had won the 2004 Gödel Prize on the topological structure<sup>2</sup> of asynchronous computability and showed that the solvability of the renaming problem in asynchronous systems depends on whether the number of processes is a power of a prime number. Castañeda won the Best Student Paper Award at PODC 2008, and the paper was included in the ACM list of Notable Computing Books and Articles of 2012. The Babel file system is a software defined, massive, scalable and fault-tolerant distributed storage system developed by Ricardo Marcelín Jiménez (UAM, Mexico). Babel is a middleware for managing replicated databases, based on Paxos, and includes several innovations. Early research in Brazil on dependable distributed systems started perhaps with Rogerio Drummond (UNICAMP), who with Ozalp Babaoglu (then in the U.S.) published seminal work on clock synchronization, as well as the “Streets of Byzantium” paper on implementing reliable broadcasts in distributed systems with broadcast networks.

It was a Brazilian researcher, Joni Fraga (UFSC, Brazil), who coined the term “intrusion tolerance” in 1985. The term has been adopted worldwide and became truly popular years later with the growth of the Internet and related security problems. In coopera-

tion with his previous Ph.D. students Alysson Bessani (Portugal), Eduardo Alchieri (UnB, Brazil), and the late Lau Cheuk-Lung, they proposed relevant algorithms and tools for distributed systems that tolerate Byzantine faults. DepSpace is a system to improve the dependability of tuple spaces. Connectivity requirements for solving Byzantine consensus with unknown participants was studied with Fabiola Greve (UFBA, Brazil).

The well-known BFT-Smart system for state machine replication was proposed by Alchieri together with Bessani. With Fernando Pedone (Switzerland) they have investigated strategies to boost the concurrency of parallel state machine replication.

Raimundo Macêdo (UFBA, Brazil) proposed the concept of causal blocks to represent group message ordering. Based on this concept, with Paul Ezhilchelvan (England) and Santosh Shrivastava (England), they proposed the early and well-cited fault-tolerant, general-purpose Newtop protocol for partitionable and overlapping process groups, as well as pioneering mechanisms for flow control in group communication. The investigation of mobile process groups with virtual synchrony was an innovative contribution with Flavio Assis-Silva (Brazil). Distinct prominent aspects of consensus have been investigated: such as mobile systems with Michel Hurfin (France) and Nadjib Badache (Algeria); general agreement framework with Hurfin, Raynal, and Frederic Tronel (France); and adaptive message patterns with Hurfin, Mostefaoui, and Raynal. Macêdo addressed adaptive failure detection in asynchronous systems and the use of neural networks and Simple Network Management Protocol for adaptive detectors with his supervised student Fabio Ramon (Brazil).

Macêdo proposed alternative hybrid distributed system models that welded both synchronous and asynchronous assumptions under the same framework, initially with his Ph.D. student Sérgio Gorender (Brazil), and later with Raynal and Gorender. Group communication and simulation tools in this hybrid model have been developed in the Ph.D. thesis of Allan Freitas (Brazil). Macêdo also proposed the partitioned-syn-

chronous model with Gorender, and some problems have been addressed under this model like failure detection, mutual-exclusion, and with his supervised students Marcos Ramos, Anne Blagojevic, and Wellington Silva, Byzantine failures.

Self-manageable distributed system protocols inspired by the feedback control theory, where protocol's objectives can be modified and controlled at runtime, were also innovative contributions by Macêdo and his supervised students. The Ph.D. theses of Alirio Sá (Brazil), Freitas, and Sandro Andrade (Brazil) have applied these principles to QoS-based self-configuring failure detectors, self-manageable group communication, adaptive Byzantine replication, and self-adaptive software architecture design. Finally, an initial effort to secure IoT-based cyber-physical human systems against collaborative attacks has been undertaken with Sathish Kumar, Bharat Bhargava, and Ganapathy Mani (U.S.).

Multiple other groups have worked on diverse aspects of dependable distributed systems. Improving the precision of failure detectors using time series to predict communication delays was proposed by Ingrid Jansch-Porto (UFRGS, Brazil) with Raul Ceretta (UFSM, Brazil). Ceretta has led multiple projects on security and has had a partnership with the Universidad de Paraguay for more than 10 years. The Impact Failure Detector, which takes into account both process relevance and confidence in the system to assess the state of monitored processes, was proposed by Cláudio Geyer (UFRGS, Brazil) with Luciana Arantes (France) and Anubis Rossetto (IFSUL, Brazil). In cooperation with Pedone, Fernando Dotti (PUCRS, Brazil) has worked on several aspects of parallel state machine replication, also in cooperation with Odorico Mendizabal (UFSC). Among the several relevant results of this fruitful cooperation is the Byzantine fault-tolerant atomic multicast strategy proposed with Bessani. The dependability of streaming systems has been investigated by Andrey Britto (UFCG), with Christof Fetzer (Germany). Britto's main focus is on security. Total order broadcast and consensus have been investigated by Luiz Buzato (UNICAMP). Together with Islene

Garcia (UNICAMP) they have proposed relevant strategies for checkpoint and rollback.


Dynamic distributed systems with unknown participants have been investigated by Greve, including failure detection and eventual leader election in evolving mobile networks in cooperation with Arantes. A consensus algorithm for systems with unknown participants in shared memory was developed with Catia Khouri (IFBA, Brazil) and Sébastien Tixeuil (France). Also, with Tixeuil, Greve has investigated the knowledge connectivity versus synchrony requirements for consensus in unknown networks. A solution to the group priority inversion problem in the timed asynchronous model was proposed by Greve in cooperation with Francisco "Fubica" Brasileiro (UFCG, Brazil), Emmanuelle Anceaume (France), and Hurfin. Earlier, Greve and Brasileiro, with Mostefaoui and Raynal, in a seminal work proposed consensus in one communication step. Brasileiro, with Livia Sampaio (UFCG), proposed an adaptive process ordering module to improve the performance of adaptive indulgent consensus protocols. The implementation of fail-silent nodes for distributed systems was an early work of Brasileiro with Neil Spears (U.K.).

Working on the frontier of distributed systems and networking, Elias P. Duarte Jr. (UFPR, Brazil) has proposed with Luis Bona (UFPR) the VCube virtual topology for distributed systems. VCube is a hypercube when all processes are correct, but as processes fail and recover the structure reorganizes itself, keeping several logarithmic properties. The topology was first introduced in the context of distributed diagnosis with Takashi Nanya (Japan). Multiple distributed algorithms have been proposed for the VCube, including reliable broadcast and distributed mutual exclusion with Luiz A. Rodrigues (UNIOESTE, Brazil) and Arantes, and a publish-subscribe algorithm, with Pierre Sens.

Duarte has also worked on the diagnosis of dynamic partitionable general topology networks and comparison-based diagnosis. Results include a survey covering 30 years of research in this field, as well as a nearly optimal algorithm for general topologies. The search for a fault-tolerant

routing strategy for the Internet led to the development of connectivity numbers, which are centrality metrics that reflect network node connectivity, proposed with Jaime Cohen (UEPG, Brazil). The two also proposed parallel algorithms for cut trees, a very relevant combinatorial data structure. Recent work includes a strategy to improve the dependability of cloud systems based on replicating the cloud manager itself. With Rogerio Turchetti (UFSM) and Edson Camargo (UTFPR), they have investigated the dependability of programmable/virtualized networks, including the usage of network function virtualization technologies to implement in-network distributed services, including failure detectors, reliable and ordered broadcast, and consensus.

## Conclusion

To conclude, we must remark that the article is far from exhaustive, not only in terms of results of the research groups mentioned, but also because there are many other groups that could not be discussed due to space restrictions. In particular, security is a dependability attribute and there is a very large number of groups working in that field. Most of the research groups presented here are at universities that are forming a good number of young researchers who are very enthusiastic about dependable computing research. The future looks bright! 

## References

1. Avizienis, A. et al. Basic concepts and taxonomy of dependable and secure computing. *IEEE Trans. Dependable and Secure Computing* 1.1 (2004), 11–33.
2. Herlihy, M., Kozlov D.N. and Rajsbaum S. *Distributed Computing Through Combinatorial Topology*. Morgan Kaufmann, 2013.
3. Pradhan, D.K. (Ed.). *Fault-Tolerant Computer System Design*. Prentice-Hall, 1996.
4. von Neumann, J. Probabilistic logics and synthesis of reliable organisms from unreliable components. *Automata Studies*. C. Shannon, J. McCarthy, J. (Eds.). Princeton University Press, Princeton, NJ, 1956, 43–98.

**Elias P. Duarte Jr.** is a professor in the Department of Informatics at Federal University of Paraná, Curitiba, Brazil.

**Raimundo J. A. Macêdo** is a professor in the Computer Science Department at Federal University of Bahia, Salvador, Brazil.

**Eliane Martins** is a collaborating professor in the Institute of Computing at the State University of Campinas, Campinas, Brazil.

**Sergio Rajsbaum** is a professor at the Instituto de Matemáticas at the Universidad Nacional Autónoma de México, in Mexico City, México.



BY MARCOS KIWI, YOSHIHARU KOHAYAKAWA,  
SERGIO RAJSBAUM, FRANCISCO RODRÍGUEZ-HENRÍQUEZ,  
JAYME LUIZ SZWARCFITER, AND ALFREDO VIOLA

# A Perspective on Theoretical Computer Science in Latin America

THEORETICAL COMPUTER SCIENCE is everywhere, for TCS is concerned with the foundations of computing and computing *is* everywhere! In the last three decades, a vibrant Latin American TCS community has emerged: here, we describe and celebrate some of its many noteworthy achievements.

Computer science became a distinct academic discipline in the 1950s and early 1960s. The first CS department in the U.S. was formed in 1962, and by the 1970s virtually every university in the U.S. had one. In contrast, by the late 1970s, just a handful of Latin American universities were actively conducting research in the area. Several CS departments were eventually established during the late 1980s. Often, theoreticians played a decisive role in the foundation of these departments.

One key catalyst in articulating collaborations among the few but growing number of enthusiastic theoreticians who were active in the international academic arena was the foundation of regional conferences. The first one was LATIN in 1992 followed by LAGOS and Latincrypt as well as other more specialized or local meetings (see the sidebars in this article for details). These conferences have fostered regional and international collaboration and helped consolidate TCS research groups in Argentina, Brazil, Chile, Mexico, and Uruguay, and their impact is felt in other Latin American countries.

In this article, we briefly discuss some of the most notorious research topics in TCS in Latin America. Our perspective is inspired by the research scope of LATIN, LAGOS, and Latincrypt; we have grouped them into nine topics.

## Automata Theory and Networks

One of the main instigators of interest and research in TCS in Brazil was Imre Simon, whose work in automata theory was very influential (for details see Pin<sup>1</sup>). Owing to Simon's research, a semiring and some of its variants were dubbed tropical semirings, and the name has endured as they became fashionable in algebraic geometry. São Paulo's theory group grew in several directions under Simon's leadership. In 1992, he launched the first Latin American theory conference (LATIN), thus fostering the emergence of a vibrant regional community. In Chile, a somewhat similar story took place. Eric Goles returned from Grenoble in the early 1980s and continued his Ph.D. thesis work on dynamics of cellular automata via discrete Lyapunov functions. He mentored several of the first Chilean TCS researchers, who, together with his more recent students, are active throughout several institutions in Chile, working in graph theory, distributed computing, Boolean networks, and so forth. Not surpris-



ingly, the second edition of LATIN took place in Chile in 1995 and was co-chaired by Goles.

### Graph Theory

Another main source of TCS development in the region has been graph theory, which started about 50 years ago. The pioneering work by the late Victor Neumann-Lara and Jayme Szwarcfiter (UFRJ, Rio de Janeiro) has had great influence in Latin America.

The research interests and achievements of Latin American graph theorists are too broad to be highlighted briefly. Undoubtedly, the beautiful Lucchesi-Younger minimax theorem on directed cuts,<sup>6</sup> by Cláudio L. Lucchesi (Unicamp, Campinas), is one of the best-known graph theory results by a Latin American. In addition to

Unicamp, in São Paulo, research in graph theory has a long tradition at USP.<sup>4</sup> Much of the graph theory in Argentina and Brazil can be traced back to UFRJ, where Szwarcfiter works. In Rio de Janeiro (at UERJ, UFF, UFRJ) the main areas are graph convexity, graph classes, and graph algorithms.<sup>2,5</sup> Furthermore, there are important researchers at UFC (Fortaleza) and active groups at UFABC (São Paulo), UFMG (Belo Horizonte), UFMS (Campo Grande), and UFPE (Recife). In Argentina, significant contributions have been made on intersection graphs and graph complexity, both at UBA (Buenos Aires) and UNLP (La Plata). The collaboration of Flavia Bonomo (UBA) and Maya Stein (UCh, Santiago) has recently resulted in a noteworthy success in classical algorithmic graph

theory, namely, in graph coloring.<sup>3</sup> Graph theory is now a major area of research in Chile, with Martín Matamala (UCh) as one of its senior leaders. There are many strong active researchers in Mexico City working in graph theory, among others at UAM, UNAM, CINVESTAV and ITAM, and in several other cities, especially Gelasio Salazar at UASLP (San Luis Potosi).

### Pattern Matching and Information Retrieval

Latin America has a strong tradition in research on string searching and information retrieval (IR), most of which can be traced back to the group led by Gaston Gonnet (from Uruguay) at Waterloo. Two of Gonnet's Ph.D. students, Nivio Ziviani (UFMG, Belo Horizonte) and Ricardo Baeza-Yates



One key catalyst in articulating collaborations among the few but growing number of enthusiastic theoreticians who were active in the international academic arena was the foundation of regional conferences.

## Main Meetings

**LATIN.** The *Latin American Theoretical INformatics Symposium* started in 1992. It is currently in its 14th edition. Since 1998 it has been held every two years. More than 1,900 papers have been submitted to these meetings and 719 have been accepted. The authors of the published papers come from nearly 50 countries and about 1/6 of these papers have an author with a Latin American affiliation and 1/12 have all their authors affiliated to Latin American institutions. Many other articles are written by the Latin American diaspora working mostly in Europe and North America. In terms of worldwide reach measured by origin of accepted articles, the top 10 list is dominated by the U.S., France, and Germany, and is complemented by Canada, Brazil, Chile, Italy, U.K., Israel, and Switzerland. Among the strong research areas for which LATIN is a thoroughfare are algorithms, computational complexity, data structures, pattern matching, and random structures.

Although effectively a four-full-day meeting, LATIN plays out over five days. This, coupled with its single session format gives ample time for interaction among attendees. Another distinctive aspect of the conference is its relatively large lineup of invited speakers that includes ACM Fellows, ACM A.M. Turing Award and Nevanlinna Prize recipients. Although LATIN keeps its Latin American nature, it is a meeting that reaches the world.

**Latincrypt.** The *International Conference on Cryptology and Information Security in Latin America* and the *Advanced School on Cryptology and Information Security in Latin America* (ASCRypt) have been held several times since 2010. Latincrypt enjoys the "In cooperation with" status granted by the International Association for Cryptologic Research (IACR). Both events are the leading cryptographic periodic meetings in Latin America. One distinctive aspect of Latincrypt is its outstanding list of invited speakers, which includes several IACR Fellows and ACM A.M. Turing Award recipients. Furthermore, during all of its editions, ASCRypt has served more than 1,000 students from all over the world, most of them from Latin America.

Over 320 papers have been submitted to Latincrypt and 113 have been accepted for publication. Among the latter papers, 1/5 have at least one author while 1/8 have all authors affiliated to a Latin American institution. The authors of the published papers come from nearly 30 countries and many of them, while not working in Latin America, are originally from the region.

**LAGOS.** The *Latin American Algorithms, Graphs and Optimization Symposium* has been held on 10 occasions, all but once in Latin America. The meeting per se arose in 2005 as the merger of two regional events that were taking place since 2001. Its proceedings were first published in *Electronic Notes in Discrete Mathematics* and later in *Electronic Notes in Theoretical Computer Science*.

The average number of submitted papers has been approximately 130. The average acceptance rate is approximately 45%. Each edition of LAGOS attracts authors from approximately 20 distinct countries. Outside Latin America, the largest number of submissions and participants are affiliated with French institutions.

(UCh, Santiago) started a very fruitful and productive collaboration in the 1990s. They co-founded, in 1993, the *International Symposium on String Processing and Information Retrieval* and pioneered the use of a novel technique called bit-parallelism in string matching algorithms. Gonzalo Navarro (UCh, Santiago), a Ph.D. student of Baeza-Yates, extended and implemented the technique, publishing a well-received book,<sup>9</sup> and developing the public software *nrgrep*. In the late 1990s, the groups at UFMG and Universidad de Chile developed a new area: direct search on compressed natural language text.<sup>10</sup> Those developments were applied in novel Web search engines devised in Chile and Brazil, which were eventually instrumental in the

interest of Yahoo! and Google to settle in Chile and Brazil, respectively. In the mid-1990s, Berthier Ribeiro-Neto joined UFMG, bringing his experience in core IR and ranking to the group. In 1999, with Baeza-Yates, he published the book *Modern Information Retrieval*.<sup>7</sup> This is one of the most-cited publications in the history of IR, with close to 19K citations at the time of this writing, according to Google Scholar Citations. Since 2000, Navarro's research has focused on compressed data structures,<sup>8</sup> which is described in the article by Arroyuelo et al. on p. 64.

A number of young researchers in the area now populate various regional universities, especially in Chile and Brazil but also in Colombia, Ecuador, and Mexico.

## Algorithms

Algorithms research in Latin America is particularly successful in the areas of approximation algorithms, online algorithms, and algorithmic game theory. Excellent groups are found in Chile and Brazil. The community in Chile has been growing steadily and today the main groups are based in UCh, PUC-Chile, USACH, and UOH. Recent outstanding results are summarized next. Andreas Wiese (UCh, Santiago) obtained a  $(1 + \epsilon)$ -approximation algorithm for finding a maximum weight independent set of polygons in quasi-polynomial time,<sup>11</sup> José Correa (UCh) resolved an open problem from the 1980s related to the IID prophet inequality,<sup>13</sup> while José Soto (UCh) and Victor Verdugo (UOH, Rancagua) proposed the best current algorithms for the secretary problem on some classes of matroids.<sup>17</sup> In Brazil, the main groups are at PUC-Rio, Unicamp, and USP. At PUC-Rio, algorithms research mainly focuses on algorithms under uncertainty and learning. Marco Molinaro (PUC-Rio, Rio de Janeiro) has been exploring connections between online and stochastic problems and online learning.<sup>12</sup> Eduardo Laber (PUC-Rio) has been working on decision tree problems and its connections with machine learning.<sup>14</sup> The groups at Unicamp and USP collaborate regularly. For instance, Flávio Miyazawa (Unicamp, Campinas) and Yoshiko Wakabayashi (USP, São Paulo) have a long history of collaboration on approximation algorithms for geometric packing problems,<sup>16</sup> while Cristina Fernandes (USP), Flávio Miyazawa, Luis Meira, and Leilton Pedrosa (Unicamp) have developed a systematic technique to bound factor-revealing linear programs and used it to obtain approximation results for facility location problems.<sup>15</sup> Also in São Paulo, Marcel de Carli Silva (USP) and Cristiane Sato (UFABC, Santo André) have worked on spectral sparsification algorithms.

## Distributed Algorithms

In Mexico, there is an active research group in distributed algorithms since the early 1990s, started by Sergio Rajsbaum, which later incorporated Armando Castañeda, both at UNAM

(Mexico City). They are internationally recognized as some of the main experts on the topological approach to distributed computing. This perspective was born in 1993 when Maurice Herlihy and Nir Shavit and others, uncovered a deep connection with algebraic topology, showing that communication among unreliable concurrent processes is actually deforming a geometric representation of the possible inputs to the system, and the topological properties in turn determine computability and complexity of the corresponding distributed algorithms. The long-term collaboration since the early 1990s especially with Herlihy resulted in the 2013 book<sup>22</sup> and the organization in Mexico of the 10<sup>th</sup> Geometric and Topological Methods in Computer Science conference. Close international collaborations have been maintained, especially with the U.S., France, and Israel. Some research highlights of the topological approach demonstrate its interaction with formal methods, with network algorithms,<sup>20</sup> with robot algorithms,<sup>18</sup> and with

epistemic logic.<sup>21</sup> In Chile, distributed computing research is done by Iván Rapaport (UCh, Santiago), Pedro Montealegre (UAI, Santiago) and Karol Suchan (UDP, Santiago), who have worked on communication complexity,<sup>19</sup> cellular automata, routing, and distributed property testing.

## Combinatorics and Random Structures

There is solid collaboration among researchers from Argentina and Uruguay in analytic combinatorics and dynamical analysis of algorithms. An illustration of the research done in this area is Alfredo Viola's complete distributional analysis of linear probing,<sup>26</sup> generalizing Donald Knuth's work from 1962, considered to be the origin of analysis of algorithms.

Asymptotic and probabilistic combinatorics are important topics of study in Brazil and Chile. The research in this area in Chile is led by Marcos Kiwi and Maya Stein (UCh, Santiago) and Hiệp Hàn (USACH, Santiago). Among noteworthy contributions are the proof of an approximate

## Other Meetings

**SPIRE.** The *International Symposium on String Processing and Information Retrieval*, held annually since 1993 (initially under a different name), alternates its venues between Latin America and the rest of the world. While it focuses on string processing, it also features relevant articles on information retrieval and computational biology, particularly on algorithmic and efficiency aspects. SPIRE is a key meeting point between the Latin American and the international communities working around its topics of interest.

**LAWCG.** The *Latin American Workshop on Cliques of Graphs* is a biennial meeting originally focused on clique graphs (the intersection graphs of the maximal cliques of a graph) a research topic that mostly originated in Latin America. However, it soon broadened its spectrum and became a forum for research in graph theory. The workshop attracts about 150 participants from Argentina, Brazil, Chile and Mexico, and is being held alternately in these countries.

**ACCOTA.** The biennial *International Workshop on Combinatorial and Computational Aspects of Optimization, Topology and Algebra* is a meeting at the crossroads of TCS, combinatorics, graph theory, geometry, topology, and algebra. Started in 1996, it has always taken place in Mexico, exerting great influence on the mathematics community both within and beyond Mexican borders. ACCOTA has benefited from the often regular participation of researchers of the highest level, including some of the best living graph theorists and computer scientists.

**Schools.** Many high-level schools are held periodically in the region. One of the longest running and more prestigious is the *Discrete Mathematics Summer School*, launched in 2004 by researchers at Universidad de Chile. It takes place annually on a scenic hilltop overlooking the port of Valparaiso, Chile. It has become a lively and consolidated event that attracts students from all over the region and beyond. Since 2008, the biennial *Encuentro Colombiano de Combinatoria* brings together the world's top researchers in algebraic and geometric combinatorics and undergraduate and graduate students from all over Latin America and the world.



# Acronyms of Universities and Research Institutes

- ▶ CINVESTAV: Centro de Investigación y de Estudios Avanzados del IPN, Mexico
- ▶ IMPA: Instituto de Matemática Pura e Aplicada, Brazil
- ▶ IPN: Instituto Politécnico Nacional, Mexico
- ▶ ITAM: Instituto Tecnológico Autónomo de México, Mexico
- ▶ PUC-Chile: Pontificia Universidad Católica de Chile, Chile
- ▶ PUC-Rio: Pontificia Universidade Católica do Rio de Janeiro, Brazil
- ▶ UAI: Universidad Adolfo Ibáñez, Chile
- ▶ UAM: Universidad Autónoma Metropolitana, Mexico
- ▶ UASLP: Universidad Autónoma de San Luis Potosí, Mexico
- ▶ UBA: Universidad de Buenos Aires, Argentina
- ▶ UCh: Universidad de Chile, Chile
- ▶ UdeLaR: Universidad de la República, Uruguay
- ▶ UDP: Universidad Diego Portales, Chile
- ▶ UERJ: Universidade do Estado do Rio de Janeiro, Brazil
- ▶ UFABC: Universidade Federal do ABC, Brazil
- ▶ UFC: Universidade Federal do Ceará, Brazil
- ▶ UFF: Universidade Federal Fluminense, Brazil
- ▶ UFMG: Universidade Federal de Minas Gerais, Brazil
- ▶ UFMS: Universidade Federal do Mato Grosso do Sul, Brazil
- ▶ UFPE: Universidade Federal de Pernambuco, Brazil
- ▶ UFRJ: Universidade Federal do Rio de Janeiro, Brazil
- ▶ UFSC: Universidade Federal de Santa Catarina, Brazil
- ▶ UNAL: Universidad Nacional de Colombia, Colombia
- ▶ UNAM: Universidad Nacional Autónoma de México, Mexico
- ▶ UNGS: Universidad Nacional de General Sarmiento, Argentina
- ▶ Unicamp: Universidade Estadual de Campinas, Brazil
- ▶ UNLP: Universidad Nacional de La Plata, Argentina
- ▶ UOH: Universidad de O'Higgins, Chile
- ▶ UR: Universidad del Rosario, Colombia
- ▶ USACH: Universidad de Santiago de Chile, Chile
- ▶ USP: Universidade de São Paulo, Brazil

version of a celebrated mid-1990s conjecture<sup>25</sup> and the development of the theory of random models of complex networks.<sup>27</sup> In Brazil, Yoshiharu Kohayakawa (USP, São Paulo) and Rob Morris (IMPA, Rio de Janeiro) and their collaborators work in this area. An example of a USP/IMPA collaboration on this front is a paper on the structure of dense graphs with high chromatic number,<sup>23</sup> which was awarded the Fulkerson Prize in 2018. A striking regional research success is the discovery of “hypergraph con-

tainers” by Morris and co-authors.<sup>24</sup> These objects were in fact independently and simultaneously born in two places: in Rio de Janeiro and at Cambridge, U.K., and the authors were jointly awarded the Pólya Prize in Applied Combinatorics in 2016 for their far-reaching discovery.

## Computational Geometry

Computational geometry has been present in Latin America for at least 20 years, mainly in Mexico and more recently in Chile. In Mexico, the

senior researcher in this area is Jorge Urrutia (UNAM, Mexico City) who works with David Flores-Peñaloza and Adriana Ramírez-Vigueras (UNAM), Ruy Fabila-Monroy and Dolores Lara (CINVESTAV, Mexico City) and Marco Heredia (UAM, Mexico City). This group works in several areas such as those surveyed by Urrutia,<sup>32</sup> with well-known results in routing in ad hoc wireless networks, geometric graphs on colored point sets,<sup>30</sup> orthogonal convex hulls, and discrete geometry.<sup>28</sup> The group maintains strong international collaborations, especially with research groups in Austria, Spain, Canada, and Japan, and in particular in Chile with Pablo Pérez-Lantero (USACH, Santiago). An independent research group that includes Jeremy Barbay (UCh, Santiago) has developed instance optimal algorithms for geometric problems such as computing the convex hull of a point set.<sup>29</sup>

We close by mentioning an interesting and early chapter of computational geometry involving a Latin American: Jorge Stolfi (Unicamp, Campinas) developed the celebrated quad-edge data structure<sup>31</sup> as a graduate student at Stanford in the 1980s.

## Computational Algebra and Algebraic Geometry

Buenos Aires has a strong research group in computational algebra, complexity, and algebraic geometry. Their focus is on the complexity of natural algebraic problems, such as solving systems of polynomial equations and related enumeration questions. Joos Heintz (UBA, Buenos Aires) has been a leader in these areas for decades, substantially contributing to the complexity of equation solving.<sup>33</sup> He co-founded the international meetings MEGA and TERA, and has strong collaborations with Spain, France, and Germany. The algebraic geometer Guillermo Matera (UBA, UNGS, Buenos Aires) works on computational and enumeration problems.<sup>34</sup> Eda Cesaratto’s (UNGS) work focuses on computational number theory. In Colombia, there is a very active group of young researchers who have made important contributions in algebraic and geometric combinatorics, particularly in matroid theory, polytopes, and tropical geometry.

## Cryptography

The main Latin American cryptographic research groups are concentrated in Brazil, Chile, Colombia, Mexico, and Uruguay.

The pioneering work on pairing-based cryptography of Paulo Barreto (USP, São Paulo) culminated with the discovery of the Barreto-Naehrig elliptic curves.<sup>35</sup> Several elegant algorithmic improvements introduced by Diego Aranha, Ricardo Dahab and Julio López (Unicamp, Campinas) produced some of the fastest and most compact software implementations of elliptic curve cryptography.<sup>36,38</sup> Ricardo Custódio (UFSC, Florianópolis) has contributed to digital signature systems. Jeroen van der Graaf (UFMG, Belo Horizonte) has worked in cryptographic protocols and electronic voting. Nicolas Thériault (USACH, Santiago) has worked on index calculus cryptanalysis attacks. Recently, Alejandro Hevia (UCH, Santiago) and collaborators developed the randomness beacon. John Baena and Daniel Cabarcas (UNAL, Bogotá) and Valérie Gauthier (UR, Bogotá) are junior researchers whose main interests lie in multivariate-based and code-based cryptography. In Mexico, Nareli Cruz-Cortés (IPN, Mexico City), Francisco Rodríguez-Henríquez (CINVESTAV, Mexico City), and their collaborators, currently hold the record for computing discrete logarithms on fields of characteristic three.<sup>39</sup> Cuauhtemoc Mancillas (CINVESTAV) has designed authenticated encryption schemes. Alfredo Viola (UdelaR, Montevideo) has worked on the cryptanalysis of historical Uruguayan documents and in combinatorial constructions of Boolean cryptographic functions.

## Conclusion

We have surveyed some of the Latin American achievements in TCS focusing on the scope of LATIN, LAGOS, and Latincrypt. We look forward to the future of Latin American research in TCS, and hope that sooner than later we will not only be able to celebrate that a Turing Award winner was born in the region (as Manuel Blum), but also that she or he was educated and produced major work in the Latin American academic environment.

## Acknowledgments

The following people have kindly contributed with information and comments to this article: Federico Ardila, Marcelo Arenas, Ricardo Baeza-Yates, Flavia Bonomo-Braberman, José R. Correa, Cristina G. Fernandes, Celina M. H. de Figueiredo, Joachim von zur Gathen, Eric Goles, Claudio Gutierrez, Arnaldo Mandel, Marco Molinaro, Gonzalo Navarro, Daniel Panario, Gelasio Salazar, José A. Soto, Maya Stein, Jorge Urrutia, and Yoshiko Wakabayashi. 

### Further Reading

#### Automata and Network Theory

1. Pin, J. The influence of Imre Simon's work in the theory of automata, languages and semigroups. *Semigr. Forum* 98 (2019), 1–8.

#### Graph Theory

2. Alcon, L. et al. The complexity of clique graph recognition. *Theoret. Comput. Sc.* 410 (2009), 2072–2083.
3. Bonomo, F. et al. Three-coloring and three-list-coloring of graphs without induced paths of seven vertices. *Combinatorica* 38 (2018), 779–801.
4. Grötschel, M., Thomassen, C., and Wakabayashi, Y. Hypotractable digraphs. *J. Graph Theor.* 4, 4 (1980), 377–381.
5. Itai, A., Papadimitriou, C., and Szwarcfiter, J. Hamilton paths in grid graphs. *SIAM J. Comput.* 11 (1982), 676–686.
6. Lucchesi, C.L. and Younger, D. A minimax theorem for directed graphs. *J. Lond. Math. Soc.* 2, 17 (1978), 369–374.

#### Pattern Matching and Information Retrieval

7. Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*. ACM Press (2<sup>nd</sup> ed), 2011.
8. Navarro, G. *Compact Data Structures—A Practical Approach*. Cambridge University Press, 2016.
9. Navarro, G., and Raffinot, M. *Flexible Pattern Matching in Strings—Practical On-line Search Algorithms for Texts and Biological Sequences*. Cambridge University Press, 2002.
10. Ziviani, N., de Moura, E., Navarro, G., and Baeza-Yates, R. Compression: A key for next-generation text retrieval systems. *IEEE Computer* 33, 11 (2000), 37–44.

#### Algorithms

11. Adamszek, A., Har-Peled, S., and Wiese, A. Approximation schemes for independent set and sparse subsets of polygons. *J. ACM* 66, 4 article 29 (2019).
12. Buchbinder, N. et al. k-Servers with a smile: Online algorithms via projections. In *Proceedings of 2019 SODA* (2019), 98–116.
13. Correa, J. et al. Posted price mechanisms for a random stream of customers. In *Proceedings of 2017 EC* (2017), 169–186.
14. Cicalese, F., Laber, E. S., and Murtinho, L. New results on information theoretic clustering. In *Proceedings of 2019 ICML*, 1242–1251.
15. Fernandes, C.G. et al. A systematic approach to bound factor-revealing LPs and its application to the metric and squared metric facility location problems. *Math. Program.* 153 (2015), 655–685.
16. Miyazawa, F.K. and Wakabayashi, Y. Approximation algorithms for the orthogonal z-oriented three-dimensional packing problem. *SIAM J. Comput.* 29, 3 (2000), 1008–1029.
17. Soto, J., Turkietaub, A., and Verdugo, V. Strong algorithms for the ordinal matroid secretary problem. In *Proceedings of 2018 SODA* (2018), 715–734.

#### Distributed Algorithms

18. Alcantara, M. et al. The topology of look-compute-move robot wait-free algorithms with hard termination. *Distrib. Comput.* 32, 3 (2019), 235–255.

19. Becker, F. et al. The impact of locality in the broadcast congested clique model. *SIAM J. Discret. Math.* 34, 1 (2020), 682–700.
20. Castañeda, A. et al. A topological perspective on distributed network algorithms. In *Proceedings of 2019 SIROCCO* (2019), 3–18.
21. Goubault, E., Ledent, J., and Rajsbaum, S. To appear in *Inf. Comput.* 2020, A preliminary version of A simplicial complex model for dynamic epistemic logic to study distributed task computability. In *Proceedings of 2018 GandALF* (2018), 73–87.
22. Herlihy, M., Kozlov, D.N., and Rajsbaum, S. *Distributed Computing Through Combinatorial Topology*. Morgan Kaufmann, 2013.

#### Combinatorics and Random Structures

23. Allen, P. et al. The chromatic thresholds of graphs. *Adv. Math.* 235 (2013), 261–295.
24. Balogh, J., Morris, R., and Samotij, W. Independent sets in hypergraphs. *J. Amer. Math. Soc.* 28, 3 (2015), 669–709.
25. Hladký, J. et al. The approximate LoebL-Komlós-Sós conjecture, Parts I, II, III, and IV. *SIAM J. Discrete Math.* 31, 2 (2017), 945–1148.
26. Janson, J. and Viola, A. A unified approach to linear probing hashing with buckets. *Algorithmica* 75, 4 (2016), 724–781.
27. Kiwi, M. and Mitsche, D. Spectral gap of random hyperbolic graphs and related parameters. *Ann. Appl. Probab.* 28 (2018), 941–989.

#### Computational Geometry

28. Arocha, J.L. et al. Very colorful theorems. *Discret. Comput. Geom.* 42, 2 (2009), 142–154.
29. Afshani, P., Barbay, J., and Chan, T. M. Instance-optimal geometric algorithms. *J. ACM* 64, 1 article 3 (2017).
30. Bereg, S. et al. On balanced 4-holes in bichromatic point sets. *Comput. Geom.* 48, 3 (2015), 169–179.
31. Guibas, L. and Stolfi, J. Primitives for the manipulation of general subdivisions and the computation of Voronoi diagrams. *ACM Trans. Graph.* 4, 2 (1985), 74–123.
32. Urrutia, J. Art gallery and illumination problems. In *Handbook on Computational Geometry*, J.R. Sack and J. Urrutia, Eds., North Holland (Elsevier Science Publishers), (2000), 973–1026.

#### Computational Algebra and Algebraic Geometry

33. Bank, E. et al. Intrinsic complexity estimates in polynomial optimization. *J. Complex.* 30, 4 (2014), 430–443.
34. Cesaratto, E., von zur Gathen, J., and Matera, G. The number of reducible space curves over a finite field. *J. Number Theory* 133, 4 (2013), 1409–1434.

#### Cryptography

35. Barreto, P.S.L.M., and Naehrig, M. Pairing-friendly elliptic curves of prime order. In *Proceedings of 2005 SAC* (2005), 319–331.
36. Hernández, J. and Dahab, R. Fast multiplication on elliptic curves over GF(2<sup>m</sup>) without precomputation. In *Proceedings of 1999 CHES*, 316–327.
37. The Random UChile Project: <https://random.uchile.cl/>
38. Oliveira, L.B. et al. TinyPBC: Pairings for authenticated identity-based non-interactive key distribution in sensor networks. *Comput. Commun.* 34, 3 (2011), 485–493.
39. Wikipedia. Discrete logarithm records; [https://en.wikipedia.org/wiki/Discrete\\_logarithm\\_records](https://en.wikipedia.org/wiki/Discrete_logarithm_records)

**Marcos Kiwi** is a professor at the Universidad de Chile, Santiago, Chile.

**Yoshiharu Kohayakawa** is a professor at the Universidade de São Paulo, Brazil.

**Sergio Rajsbaum** is a professor at the Universidad Autónoma de México, Mexico City, Mexico.

**Francisco Rodríguez-Henríquez** is a professor at the CINVESTAV, Mexico City, Mexico.

**Jaime Luiz Szwarcfiter** is a professor at the Universidade Federal do Rio de Janeiro and Universidade do Estado do Rio de Janeiro, Brazil.

**Alfredo Viola** is a professor at the Universidad de la República, Montevideo, Uruguay.



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

## Emulating the efficiency of in-person conversations.

BY THOMAS A. LIMONCELLI

# Five Nonobvious Remote Work Techniques

THIS ARTICLE REVEALS five nonobvious techniques that make remote work successful at Stack Overflow.

Remote work has been part of the engineering culture at Stack Overflow from the outset. Some 80% of the engineering department works remotely. This enables the company to hire top engineers from around the world, not just from the New York City area (40% of the company worked remotely prior to the COVID-19 lockdown; 100% during the lockdown). Even employees who do not work remotely must work in ways that are remote-friendly.

For some companies, working remotely was a new thing when the COVID-19 pandemic lockdowns began. At first the problems were technical: IT departments had to ramp up VPN (virtual private network) capacity, human resources and infosec departments had to

adjust policies, and everyone struggled with microphones, cameras, and videoconferencing software.

Once those technical issues are resolved, the social issues become more apparent. How do you strike up a conversation as you used to do in the office? How do you know when it is appropriate to reach out to someone? How do you prevent loneliness and isolation?

Here are my top five favorite techniques Stack Overflow uses to make remote work successful on a social level.

### Tip #1: If Anyone is Remote, We're All Remote

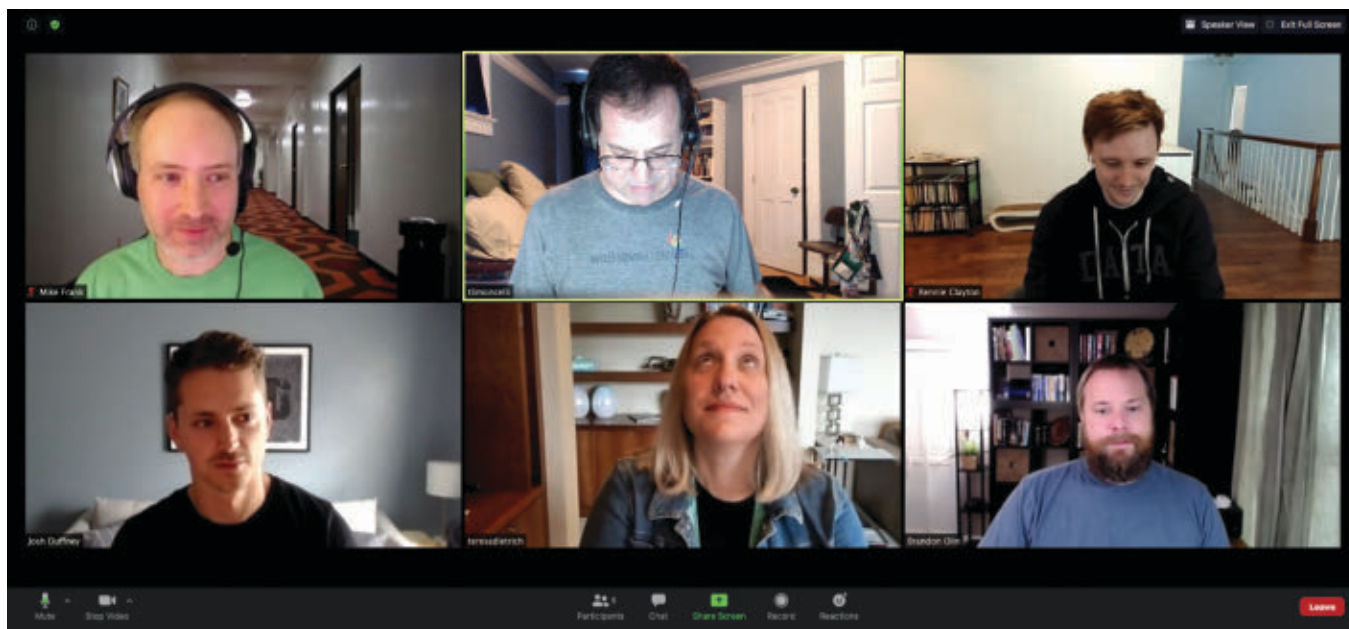
*Meetings should be either 100% in-person, or 100% remote; no mixed meetings.*

Ever been in a conference room with a bunch of people plus one person participating by phone or videoconferencing? It never works. The one remote participant can't hear the conversation, can't see what everyone else is seeing, and so on. He or she can't authentically participate.

At Stack Overflow we recognized this years ago and adopted a rule: If one person is remote, we're all remote. This means everyone who is physically present leaves the conference room, goes back to their desks, and we conduct the meeting using desktop videoconferencing.

During the COVID-19 lockdown your entire company may be remote, but this is a good policy to adopt when you return to the office.

This may not be an option for companies with open floor plans, however, where participants videoconferencing from their desks may disturb their neighbors. How can you make mixed meetings work? Where I've observed them working well required two ingredients: First, the conference-room design was meticulously refined and adjusted over time (this is rarer—and more expensive—than you would think); second, and the biggest determinant, was the degree to which all those in the meeting were aware of the remote participants. It requires a learned skill of being vigilant for clues



The author (top center) and his co-workers *not* talking to each other.

that someone is having difficulty in participating and then taking corrective action. Everyone, not just the facilitator, needs to be mindful of this.

### Tip #2: Accurate Chat Status

*Set your chat status to away when you are away. Set it to available when you are available.*

A chat system isn't just about text chatting. It fulfills the subtler purpose of indicating if a participant is present.

In meatspace you don't walk up to someone's desk and start talking. First you notice if the person is there. Talking to an empty chair is ineffective and may worry your coworkers who witness it.

Instead, you look for social clues. A shut door indicates someone needs privacy. A worker in a cubicle wearing headphones may be indicating a need for extreme focus. Someone standing up to stretch is signaling a break. Two people talking about last night's game indicates that others may join in, or may interrupt for work-related matters.

The status presented in your chat system conveys similar information. I prefer to have my status set automatically. For example, I use a feature that automatically changes my status to "In a meeting" if my Google Calendar indicates that. People get automated warnings if they try to chat to me outside of normal working hours. A "Do not disturb" feature helps control when my phone beeps for new messages or when I can simply sleep.

Since many schools and child-care facilities are closed now, parents are balancing child care and work in unexpected ways, often with highly dynamic schedules. One way we can support them is to respect status messages such as "Child care... be back in an hour" or simply "Child care," more likely to be used when there is no time to set a more detailed explanation because the kid is painting the cat with chocolate milk.

Set your status to indicate if you are away, present, present but too busy to be disturbed, and so on. Encourage your team to do the same.

Speaking of chat rooms, don't make finding the right room a guessing game. Yes, it's cute and funny to have #kittens as the chat room of the team run by a manager known for loving cats. A month later, however, nobody is laughing and it's just annoying to remember the name.

At Stack Overflow we're not super strict about names, but most rooms have prefixes such as #team-, #project-, #account-, #fun-, and so on. Need to find the SRE team? You can bet we're in #team-sre. Very little guessing. By using prefixes, not suffixes, the long list of room names sorts related names together.

### Tip #3: The Quick Chat Protocol

*Establish a low-overhead way to start a quick conversation.*

In meatspace you can ask a quick question by just blurting it out. You

are sitting near a coworker, so you can just say, "Got a second? Let's talk about the Tomato Incident." If the person is available, you just start talking about it.

Chat rooms are good for quick questions that can be explained in a few sentences. Some questions, however, are more easily explained conversationally. In such situations people often do a complicated dance in an attempt to be polite.

**Me:** Hi!

**You:** Hi.

**Me:** Got a sec?

**You:** Yeah.

**Me:** I have a question about the Tomato Incident.

**You:** OK.

**Me:** Can we chat by video?

**You:** Sure. Want me to make the room or should you?

**Me:** I'll do it.

**Me:** [link]

By the time we've negotiated when, how, and where we're going to talk, we've lost momentum.

Take all that overhead, multiply it by the number of casual conversations you have with coworkers, and it totals to a big waste of time. Such etiquette is useful when talking to someone you don't know well. For a direct coworker, however, you need a protocol with low overhead—as low as peeking over a cubicle wall.

At Stack Overflow, we follow some easy rules to avoid the extended conversations. We say in chat, for example, "Quick hangout? Tomato Incident." If available, your contact replies with a link to a chatroom. No formalities, no



small talk, no confusion over who creates the session. Otherwise, the contact responds, “Not now,” and it is up to the requester to make an appointment.

It looks like this:

**Me:** Quick chat? Tomato Incident.

**You:** [https://link\\_to\\_a\\_chat\\_room](https://link_to_a_chat_room) or

**Me:** Quick chat? Tomato Incident.

**You:** Busy right now, put something on my meal please.

When teams have prearranged to compress conversations that way, everyone benefits.

In general, we want the overhead of starting a conversation to match the type of meeting. Important meetings with many people require planning, careful scheduling, and perhaps even a rehearsal. That’s an appropriate amount of overhead. Impromptu meetings should have minimal start-up overhead. If a brief conversation requires custom-printed invitations with embossed lettering, a save-this-date card, and a return envelope, you’re doing it wrong.

The quick chat protocol has become the virtual equivalent of social customs such as visiting someone’s office, striking up a quick conversation when you pass by someone in the hallway, or spotting someone at the water cooler and approaching with a question.

One feature that enables low-overhead conversations is a video-chat system that supports permanent meeting-room URLs. This eliminates the need to pause and “create a room.” Zoom has PMIs (personal meeting IDs). As of this writing, Google Meet doesn’t have an equivalent, but you can schedule a meeting in Google Calendar in the future and use the embedded Meet URL as your PMI. (There are services that will automate this for you, as well as one open source project; <https://github.com/g3rv4/GMeet>.)

#### **Tip #4: Idling in a Videoconference Room**

*Work silently together in a videoconference room.*

Working remotely can be lonely. Chat rooms and video conferences go only so far.

At Stack Overflow many teams hang out together in a videoconference even when not having a meeting. This is in addition to their regular text chat

room. People just stay connected to the video chat, often with audio muted, and silently work independently. They unmute for quick questions or to consult on an idea. People drop out if they have another meeting, need to be alone, or need to focus.

Think of this as emulating the physical world where many people work in the same room or work in an open floor plan. A manager who consistently hangs out in the videoconference is simulating an open-door policy, signaling permission to go in and talk. A friend who works in child care calls this “nerd parallel play.” I’m going to assume she means it as a compliment.

It may seem like a waste of bandwidth to be in a videoconference when nobody is talking, but it helps fight loneliness and builds team cohesion.

Some teams do this more than others. Some do it during specific hours of the day. The phrase “I’m idling in perma” translates to “I’m working. I’ll be in the permanent video chat room. Feel free to join and not talk.”

#### **Tip #5: Schedule Video-Chat Social Events**

*Create social events specifically for remote workers.*

Stack hosts regular social events via videoconference. During the lockdown your family might be doing this for holidays, birthdays, anniversaries, and so on. Why not do it at work too?

A coworker set up daily video chats during lunch. Volunteers have been doing weekly cooking lessons. A film club has popped up; they agree to all watch the same film during the week and then discuss it during lunch on Friday. Diversity employee resource groups and affinity groups host regular video meetups with no agenda other than to be social.

Like many startups, we normally have a “beer bash” at the end of the week. Our virtual version of this is called “the remote bev bash.” Before the lockdown this was a smaller event; now it is companywide. It is interesting to see what new concoctions people bring each week.

#### **Let’s Stop Apologizing for Normality**

While this is not an official Stack Overflow policy, I advocate that we should stop apologizing for normality.

I’ve noticed that some people spend an inordinate amount of time apologizing in videoconferences for technical problems, difficulty finding the mute button, or children running into the room unexpectedly. In the new world of remote work these events are normal. Until videoconference software works perfectly, makes the mute button easier to find, and schools reopen, these situations will just be part of everyday life. If they are normal, we shouldn’t have to apologize for them. It is a waste of time, multiplied by the number of people in the meeting.


Instead, we should simply acknowledge what happened and move on by saying something like “Thank you for waiting while I fixed my camera” or “Ah, there’s the unmute button,” or “That’s my kid who just ran by singing the Lego song, no need to applaud.”

#### **Conclusion**

The physical world has social conventions around conversations and communication that we use without even thinking. As we move to a remote-work world, we have to be more intentional to create such conventions. Developing these social norms is an ongoing commitment that outlasts initial technical details of VPN and desktop videoconference software configuration.

Companies that previously forbade remote work can no longer deny its benefits. Once the pandemic-related lockdowns are over, many people will continue working remotely. Those who return to the office will need to work in ways that are compatible with their remotely working associates.

Every company is different. The techniques discussed in this article work for Stack Overflow but might not work everywhere. Give them a try and see what works for you.

**Acknowledgment.** Thanks to my coworkers at Stack Overflow Inc. for their feedback. 

---

**Thomas A. Limoncelli** is the SRE manager at Stack Overflow Inc. in New York City. His books include *The Practice of System and Network Administration*, *The Practice of Cloud System Administration*, and *Time Management for System Administrators*. He blogs at [EverythingSysadmin.com](http://EverythingSysadmin.com) and tweets at [@YesThatTom](https://twitter.com/YesThatTom).

Copyright held by author/owner.  
Publication rights licensed to ACM.

---

## Data kept outside SQL has different characteristics from data kept inside.

---

BY PAT HELLAND

---

# Data on the Outside versus Data on the Inside

RECENTLY, THERE HAS been a lot of interest in services. These can be microservices or just services. In each case, the service provides a function with its own code and data and operates independently of partners. This article argues that there are a number of seminal differences between data encapsulated

inside a service and data sent into the space outside of the service boundary.

SQL data is encapsulated within a service to ensure it is protected by application code. When sending data across services, it is outside that trust boundary.

The first question this article asks is what trust means to a service and its encapsulated data. This is answered by looking at transactions and boundaries, data kept inside versus data kept outside of services. Also, to be considered is how services compose using operators (requesting stuff) and operands (refining those requests). Then the article looks at time and service boundar-

ies. When data in a database is unlocked, it impacts notions of time. This leads to an examination of the use of immutability in the composition of services with messages, schema, and data flowing between these boundaries.

The article then looks at data on the outside of these trust boundaries called services. How do you structure that data so it is meaningful across both space and time as it flows in a world not inside a service? What about data inside a service? How does it relate to stuff coming in and out?

Finally, the characteristics of SQL and JavaScript Object Notation (JSON),



and other semi-structured representations, are considered. What are their strengths and weaknesses? Why do the solutions seem to use both of them for part of the job?

### Essential Services

Services are essential to building large applications today. While there are many examples of large enterprise solutions that leverage services, the industry is still learning about services as a design paradigm. This section describes how the term *service* is used and introduces the notions of data residing inside services and outside services.


**Services.** Big and complex systems are typically collections of independent and autonomous services. Each service consists of a chunk of code and data that is private to that service. Services are different from the classic application living in application silos, in that services are primarily designed to interact with other services via messaging. Indeed, that interaction, its data, and how it all works is an interesting topic.

Services communicate with each other exclusively through messages. No knowledge of the partner service is shared other than the message formats and the sequences of the expected messages. It is explicitly allowed (and, indeed, expected) that the partner service may be implemented with heterogeneous technology at all levels of the stack including hardware, operating system, database, middleware, programming language, and/or application vendor or implementation team.


The essence of a service lies in its independence and how it encapsulates (and protects) its data.

**Bounding trust via encapsulation.** Services interact with a collection of messages whose formats (schema) and business semantics are well defined. Each service will do only limited things for its partner services based upon well-defined messages. The act of defining a limited set of behaviors provides a firm encapsulation of the service. An important part of trust is limiting the things you'll do for outsiders.

To interact with a service, you must follow its rules and constraints. Each message you send fits a prescribed role. The only way to interact with data in another service is through its rules



**The essence of a service lies in its independence and how it encapsulates (and protects) its data.**



and business logic. Data is, in general, never allowed out of a service unless it is processed by application logic.

For example, when using your bank's ATM, you expect to have only a few supported operations such as withdrawal, deposit, among others. Banks do not allow direct access to database connections via ATMs. The only way to change the bank's database is through the bank's application logic in the ATM and the back-end system. This is how a service protects its data.

**Encapsulating both changes and reads.** Services encapsulate changes to their data with their application logic. The app logic ensures the integrity of the service's data and work. Only the service's trusted application logic can change the data.

Services encapsulate access to read their data. This controls the privacy of what is exported. While this autonomy is powerful, it can also cause some challenges.

Before your business separated its work into independent services, all of its data was in a big database. Now you have a bunch of services, and they have a bunch of databases running on a bunch of computers with a bunch of different operating systems. This is awesome for the independent development, support, and evolution of the different services, but it's a royal hassle when you want to do analytics across all your data.

Frequently each service will choose to export carefully sanitized subsets of data for consumption by partner services. Of course, this requires some work ensuring proper authorization to see this data (as well as authenticating the curious service). Still, the ability to sanitize and control the data being exposed is crucial.

**Trust and transactions.** Participating in an ACID (atomic, consistent, isolated, durable) transaction means one system can be locked up waiting for another system to decide to commit or abort the transaction. If you are stuck holding locks waiting for another system, that can really cause trouble for your availability. With rare exceptions, services don't trust other services like that.

In the late 1990s, there were efforts to formalize standards for transaction coordination across trust boundaries.

Fortunately, these standards died a horrible death.

### Data inside and outside services.

The premise of this article is that data residing inside a service is different in many essential ways from data residing outside or between services:

- ▶ Data on the inside refers to the encapsulated private data contained within the service itself. As a sweeping statement, this is the data that has always been considered “normal”—at least in your database class in college. The classic data contained in a SQL database and manipulated by a typical application is inside data.

- ▶ Data on the outside refers to the information that flows between these independent services. This includes messages, files, and events. It’s not your classic SQL data.

**Operators and operands.** Messages flowing between services contain *operators*, which correspond to the intended purpose of the message. Frequently the operator reflects a business function in the domain of the service. For example, a service implementing a banking application may have operators in its messages for deposits, withdrawals, and other banking functions. Sometimes operators reflect more mundane reasons for sending messages, such as “Here’s Tuesday’s price list.”

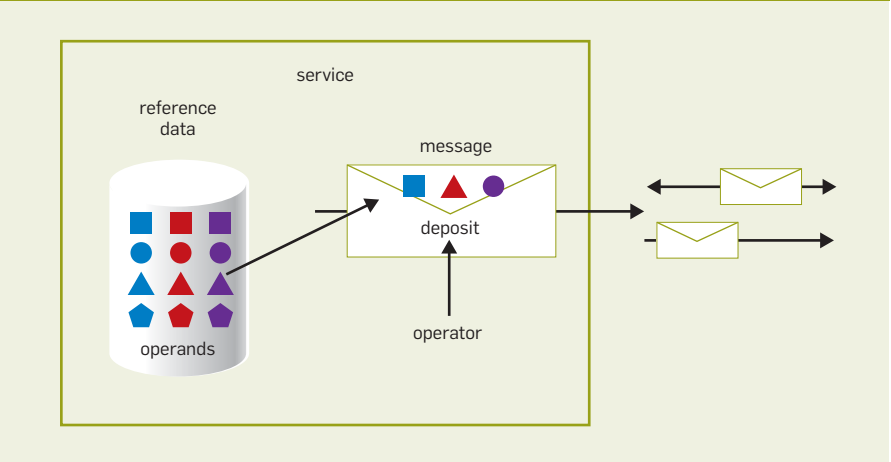
Messages may contain *operands* to the operators, shown in Figure 1. The operands are additional stuff needed by the operator message to qualify the intent of the message fully. Operands may be obtained from *reference data*, published to describe those operands. A message requesting a purchase from an e-commerce site may include product IDs, requested numbers to be purchased, expected price, customer ID, and more. This is covered in more detail later.

### Data: Then and Now

This section examines the temporal implications of not sharing ACID transactions across services and examines the nature of work inside the boundaries of an ACID transaction. This provides a crisp sense of “now” for operations against inside data.

The situation for data on the outside of the service, however, is different. The fact it is unlocked means the data is no longer in the now. Further-

Figure 1. Operands.



more, operators are requests for operations that have not yet occurred and actually live in the future (assuming they come to fruition).

Different services live in their own private temporal domains. This is an intrinsic part of using distrusting services. Trust and time carry implications of how to think about applications.

**Transactions, inside data, and now.** Transactions have been historically defined using ACID properties.<sup>1</sup> These properties reflect the semantics of the transaction. Much work has been done to describe transaction serializability, in which transactions executing on a system or set of related systems perceive their work as applied in a serial order even in the face of concurrent execution.<sup>2</sup> Transactional serializability makes you feel alone. A rephrasing of serializability is that each transaction sees all other transactions to be in one of three categories:

- ▶ Those whose work preceded this one.
- ▶ Those whose work follows this one.
- ▶ Those whose work is completely independent of this one.

This looks just like the executing transaction is all alone.

ACID transactions live in the now. As time marches forward and transactions commit, each new transaction perceives the impact of the transactions that preceded it. The executing logic of the service lives with a clear and crisp sense of now.

**Blast from the past.** Messages may contain data extracted from the local service’s database. The sending application logic may look in its belly to extract that data from its database. By the time the message leaves the service,

that data will be unlocked.

The destination service sees the message; the data on the sender’s service may be changed by subsequent transactions. It is no longer known to be the same as it was when the message was sent. The contents of a message are always from the past, never from now.

There is no simultaneity at a distance. Similar to the speed of light bounding information, by the time you see a distant object, it may have changed. Likewise, by the time you see a message, the data may have changed.

Services, transactions, and locks bound simultaneity:

- ▶ Inside a transaction, things are simultaneous.
- ▶ Simultaneity exists only inside a transaction.
- ▶ Simultaneity exists only inside a service.

All data seen from a distant service is from the “past.” By the time you see data from a distant service, it has been unlocked and may change. Each service has its own perspective. Its inside data provides its framework of “now.” Its outside data provides its framework of the “past.” My inside is not your inside, just as my outside is not your outside.

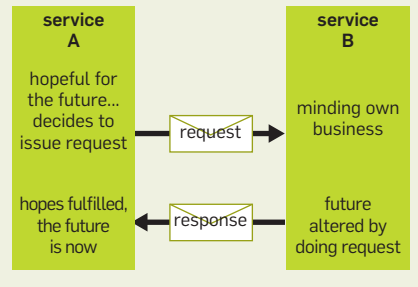
Using services rather than a single centralized database is like going from Newton’s physics to Einstein’s physics:

- ▶ Newton’s time marched forward uniformly with instant knowledge at a distance.

- ▶ Before services, distributed computing strove to make many systems look like one, with RPC (remote procedure call), two-phase commit, and so on.



Figure 2. Requests for work.



► In Einstein’s universe, everything is relative to one’s perspective.

► Within each service, there is a “now” inside, and the “past” arriving in messages.

**Hope for the future.** Messages contain operators that define requests for work from a service, shown in Figure 2. If Service A sends a message with an operator request to Service B, it is hopeful that Service B will do the requested operation.

In other words, it is hopeful for the future. If Service B complies and performs the work, that work becomes part of Service B’s future, and its state is forever changed. Once Service A receives a reply describing either success or failure of the operation, Service A’s future is changed.

**Life in the “then.”** Operands may live in either the past or the future, depending on their usage pattern. They live in the past if they have copies of unlocked information from a distant service. They live in the future if they contain proposed values that hopefully will be used if the operator is successfully completed.

Between the services, life is in the world of “then.” Operators live in the future. Operands live in either the past or the future. Life is always in the then when you are outside the confines of a service. This means that data on the outside lives in the world of then. It is past or future, but it is not now.

Each separate service has its own separate “now,” illustrated in Figure 3. The domains of transaction serializability are disjoint, and each has its own temporal environment. The only way they interact is through data on the outside, which lives in the world of then.

**Dealing with now and then.** Services must cope with making the now meet the then. Each service lives in its own now and interacts with incoming and outgoing notions of then. The application logic for the service must reconcile these.

Consider, for example, what’s involved when a business accepts an order: The business may publish daily prices, but it probably wants to accept yesterday’s prices for a while after midnight. Therefore, the service’s application logic must manually cope with the differences in prices during the overlap.

Similarly, a business that says its product “usually ships in 24 hours” must consider the following: Order processing has old information; the available inventory is deliberately fuzzy; both sides must cope with different time domains.

#### **The world is no longer flat:**

► Services with private data support more than one computer working together.

► Services and their service boundaries mean multiple trust domains and different transaction domains.

► Multiple transaction domains mean multiple time domains.

► Multiple time domains force you to cope with ambiguity to allow coexistence, cooperation, and joint work.

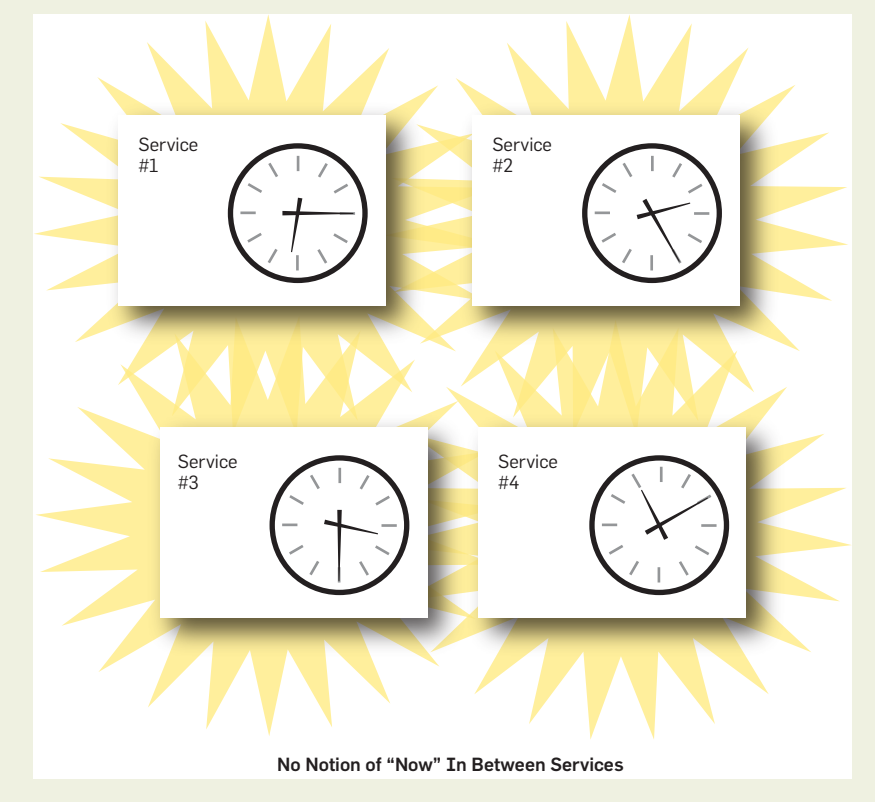
#### **Data on the Outside: Immutability**

This section discusses properties of data on the outside. First, each data item needs to be uniquely identified and have immutable contents that do not change as copies of it move around. Next, anomalies can be caused in the interpretation of data in different locations and at different times; the notion of “stable” data avoids these anomalies. The section also discusses schemas and the messages they describe. This leads to the mechanisms by which one piece of outside data can refer to another piece of data and the implications of immutability. Finally, what does outside data look like when it is being created by a collection of independent services, each living in its own temporal domain?

#### **Immutable and/or versioned data.**

Data may be immutable. Once immutable data is written and given an identifier, its contents will remain the same for that identifier. Once it is written, it cannot be changed. In many environments, the immutable data may be deleted, and the identifier will subsequently be mapped to an indication of “no present data,” but it will never return data other than the original contents. Immutable data is the same no matter when or

Figure 3. Service with different “now”s.



where it is referenced. Versioned data is immutable. If you specify a specific version of some collection of data, you will always get the same contents.


In many cases, a *version-independent identifier* is used to refer to a collection of data. An example is the *New York Times*. A new version of the newspaper is produced each day (and, indeed, because of regional editions, multiple versions are produced each day). To bind a version-independent identifier to the underlying data, it is necessary first to convert to a version-dependent identifier. For example, the request for a recent *New York Times* is converted into a request for the *New York Times* on Jan. 4, 2005, California edition.

This is a version-dependent identifier that yields the immutable contents of that region's edition of that day's paper. The contents of this edition for that day will never change no matter when or where you request it. Either the information about the contents of that specific newspaper is available or it is not. If it is available, the answer is always the same.


**Immutability, messages, and outside data.** One reality of messaging is that messages sometimes get lost. To ensure delivery, the message must be retried. It is essential that retries have the same contents. The message itself must be immutable. Once a message is sent, it cannot be unsent any more than a politician can unsay something on television. It is best to consider each message as uniquely identified, and that identifier must yield immutable contents for the message. This means the same bits are always returned for a given message.

**Stability of data.** Immutability isn't enough to ensure a lack of confusion. The interpretation of the contents of the data must be unambiguous. *Stable data* has an unambiguous and unchanging interpretation across space and time.

For example, a monthly bank statement is stable data. Its interpretation is invariant across space and time. On the other hand, the words *President Bush* had a different meaning in 2005 than they did in 1990. These words are not stable in the absence of additional qualifying data. Similarly, anything called *current* (for example, current inventory) is not stable.



**To ensure the stability of data, it is important to design for values that are unambiguous across space and time.**



To ensure the stability of data, it is important to design for values that are unambiguous across space and time. One excellent technique for the creation of stable data is the use of time-stamping and/or versioning. Another important technique is to ensure that important identifiers such as customer IDs are never reused.

**Immutable schema and immutable messages.** As discussed previously, when a message is sent, it must be immutable and stable to ensure its correct interpretation. In addition, the schema for the message must be immutable. For this reason, it is recommended that all message schemas be versioned and that each message use the version-dependent identifier of the precise definition of the message format. Alternatively, the schema can be embedded in the message. This is popular when using JSON or other semi-structured formats.

**References to data, immutability, and DAGs.** Sometimes it is essential to refer to other data. When referencing data from outside, the identifier used for the reference must specify data that is immutable.

If you find an immutable document that tells you to read today's *New York Times* to find out more details, that doesn't do you any good without more details (specifically the date and region of the paper).

As new data is generated, it may have references to complex graphs of other data items, each of which is immutable and uniquely identified. This creates a directed acyclic graph (DAG) of referenced data items. Note that this model allows for each data item to refer to its schema using simply another arc in the DAG.

Over time, independent services, each within its own temporal domain, will generate new data items blithely ignorant of the recent contributions of other services. The creation of new immutable data items that are interrelated by membership in this DAG is what gives outside data its special charm.

#### **Data on the Outside: Reference Data**

Reference data refers to a type of information that is created and/or managed by a single service and published to other services for their use. Each piece of reference data has both



a version-independent identifier and multiple versions, each of which is labeled with a version-dependent identifier. For each piece, there is exactly one publishing service.

This section discusses the publication of versions, then moves on to the various uses of reference data.

**Publishing versioned reference data.** The idea here is quite simple. A version-independent identifier is created for some data. One service is the owner of that data and periodically publishes a new version that is labeled with a version-dependent identifier. It is important that the version's identifier is known to be increasing as subsequent versions are transmitted.

When a version of the reference data is transmitted, it must be assumed to be somewhat out of date. The information is clearly from the past and not now. It is reasonable to consider these versions as snapshots.

**Uses of reference data.** There are three broad usage categories for reference data, at least so far:

- ▶ Operands contain information published by a service in anticipation that another service will submit an operator using these values.

- ▶ Historic artifacts describe what happened in the past within the confines of the sending service.

- ▶ Shared collections contain information that is held in common across a set of related services that evolves over time. One service is the custodian and manages the application of changes to a part of the collection. The other services use somewhat older versions of the information.

*Operands.* As previously discussed, messages contain operators that map to the functions provided by the service. These operators frequently require operands as additional data describing the details of the requested work. Operands are gleaned from reference data that is typically published by the service being invoked. A department store catalog, for example, is reference data used to fill out the order form. An online retailer's price list, product catalog, and shipping-cost list are operands.

*Historic artifacts.* Historic artifacts report on what happened in the past. Sometimes these snapshots of history need to be sent from one service to another.

other. Serious privacy issues can result unless proper care is exercised in the disclosure of historic artifacts from one service to another. For this reason, many times this usage pattern is seen across services that have some form of trust relationship such as quarterly results of sales, a monthly bank statement, inventory status at the end of the quarter.

*Shared collections.* The most challenging usage pattern for reference data is the shared collection. In this case, many different services need to have a recent view of some interesting data. Frequently cited examples include the employee database and the customer database. In each of these, lots of separate services want both to examine and to change the data in these collections.

Many large enterprises experience this problem writ large. Lots of different applications think they can change the customer database, and now that these applications are running on many servers, there are many replicas of the customer database (frequently with incompatible schemas). Changes made to one replica gradually percolate to the others with information loss caused by schema transformations and conflicting changes. A shared collection offers a mechanism for rationalizing the desire to have multiple updaters and allowing controlling business logic to enforce policies on the data. A shared collection has one special service that actually owns the authoritative perspective of the collection. It enforces business rules that ensure the integrity of the data. The owning service periodically publishes versions of the collection and supports incoming requests whose operators request changes.

Note that this is not optimistic concurrency control. The owning service has complete control over the changes to be made to the data. Some fields may be updatable, and others may not. Business constraints may be applied as each requested change is considered.

Consider changes to a customer's address. This is not just a simple update but complex business logic:

- ▶ You don't simply update an address. You append the new address while remembering that the old address was in effect for a range of dates.

- ▶ Changing the address may affect the tax location.

- ▶ Changing the address may affect the sales district.

- ▶ Shipments may need to be rerouted.

## Data on the Inside

As described previously, inside data is encapsulated behind the application logic of the service. This means the only way to modify the data is via the service's application logic. Sometimes a service will export a subset of its inside data for use on the outside as reference data.

This section examines the following facets of data on the inside: (1) the temporal environment in which SQL's schema definition language operates; (2) how outside data is handled as it arrives at a service; and (3) the extensibility seen in data on the outside and the challenges inherent in storing copies of that data inside in a shredded fashion to facilitate its use in relational form.

**SQL, DDL, and serializability.** SQL's Data Definition Language (DDL) is transactional. Like other operations in SQL, updates to the schema via DDL occur under the protection of a transaction and are atomically applied. These schema changes may make a significant difference in the ways that data stored within the database is interpreted.

It is essential that transactions preceding a DDL operation be based on the existing schema, and those that follow the DDL operation be based on the schema as changed by the operation. In other words, changes to the schema participate in the serializable semantics of the database.

Both SQL and DDL live in the now. Each transaction is meaningful only within the context of the schema defined by the preceding transactions. This notion of now is the temporal domain of the service consisting of the service's logic and its data contained in this database.

**Storing incoming data.** When data arrives from the outside, most services copy it inside their local SQL database. Although inside data is not, in general, immutable, most services choose to implement a convention by which they immutably retain the data. It is not uncommon to see the incoming data syntactically converted to a more convenient form for the service. This is called *shredding* (see Figure 4).

Many times, an incoming message

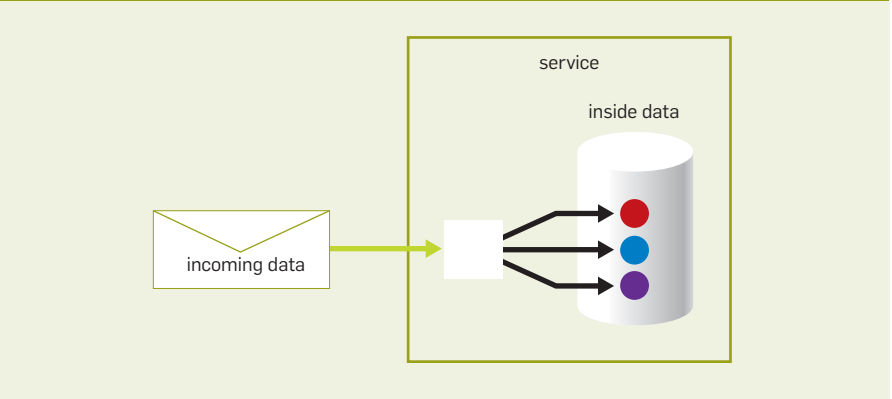
is kept as an exact binary copy for auditing and non-repudiation while the contents are converted to a form that is easier to use within the service.

**Extensibility versus shredding.** Frequently the outside data is kept in a semi-structured representation such as JSON, which has a number of wonderful qualities for this, including extensibility. JSON’s extensibility allows other services to add information to a message that was not declared in the schema for the message. Basically, the sender of the message has added stuff that you didn’t expect when the schema was defined. Extensibility is in many ways like scribbling on the margins of a paper form. It frequently gets the desired results, but there are no guarantees.

As incoming outside data is copied into the SQL database, there are advantages to shredding it. Shredding is the process of converting the hierarchical semi-structured data into a relational representation. Normalizing the incoming outside data is not a priority. Normalization is designed to eliminate or reduce update anomalies. Even though you are stuffing the data into a SQL database, you’re not going to update it. You are capturing the outside data in a fashion that’s easier to use inside SQL. Shredding is, however, of great interest for business analytics. The better the relational mapping, the better you will be able to analyze the data.

It is interesting that extensibility fights shredding. Mapping unplanned extensions to planned tables is difficult. Many times, partial shredding is performed wherein the incoming information that does comply with well-known and regular schema representations is cleanly shredded into a relational representation, and the remaining data (including extensions) is kept without shredding.

Figure 4. Shredding.



**Representations of data.** Let’s consider the characteristics of these two prominent representations of data: JSON and SQL.

**Representing data in JSON.** JSON is a standard for representing semi-structured data. It is an interchange format with a human-readable text for storing and transmitting attribute-value pairs. Sometimes a schema for the data is kept outside the JSON document. Sometimes the metadata is embedded (as attribute-value pairs) into the hierarchical structure of the document. JSON documents are frequently identified with a URL, which gives the document a unique identity and allows references to it.

It is this combination of human readability, self-describing attribute-value pairs, and global identity through URLs that make JSON so popular. Of course, its excellent and easy-to-use libraries in multiple languages help too.

**Representing data in SQL.** SQL represents relationships by values contained in cells within rows and tables. Being value-based allows it to “relate” different records to each other by their value. This is the essence of the *relational* backbone of SQL. It is precisely this value-based nature of the representation

that enables the amazing query technology that has emerged over the past few decades. SQL is clearly the leader as a representation for inside data.

**Bounded and unbounded.** Let’s contrast SQL’s value-based mechanism with JSON’s identity- and reference-based mechanism.

Relational representations must be bounded. For the value-based comparisons to work correctly, there must be both temporal and spatial bounds. Value-based comparisons are meaningful only if the contents of both records are defined within the same schema. Multiple schemas can have well-defined meaning only when they can be (and are) updated within the same temporal scope (i.e., with ACID semantics in the same database). This effectively yields a single schema. SQL is semantically based on a centrally managed single schema.

Attempts over the past 20 years to create distributed SQL databases are fine but must include a single transactional scope and a single DDL schema. If not, the semantics of relational algebra are placed under pressure. SQL only works inside a single database.

JSON is unbounded. In JSON, data is referenced using URIs (uniform resource identifiers) and not values. These URIs are universally defined and unique. Of course, every URL is a legiti-

Figure 5. Inside and outside data.

	Outside Data	Inside Data
Immutable?	Yes	No
Identity based references	Yes	No
Open schema?	Yes	No
Represent in JSON or other semi-structured fashion	Yes	No
Encapsulation useful?	No	Yes
Long-lived evolving data with evolving schema?	No	Yes
Business intelligence desirable over data?	Yes	Yes
Durable storage in SQL inside the service?	Yes	Yes

Figure 6. The dynamic duo of data representations.

	Arbitrary queries	Independent definition of shared data
SQL	Outstanding!	Impossible. SQL data definition is centralized, not independent
JSON	Problematic	Outstanding!



mate URI so they're cool, too. URIs can be used on any machine to uniquely identify the referenced data. When used with the proper discipline, this can result in the creation of DAGs of JSON documents, each of which may be created by independent services living in independent temporal (and schema) domains.

**Characteristics of inside and outside data.** Let's consider the various characteristics discussed so far for inside and outside data, as shown in Figure 5.

Immutability, identity-based references, open schema, and JSON representation apply to outside data, not to inside data. This is all part of a package deal in the form of the representation of the data, and it suits the needs of outside data well. The immutable data items can be copied throughout the network and new ones generated by any service. Indeed, the open and independent schema mechanisms allow independent definition of new formats for messages, further supporting the independence of separate services.

Next, consider encapsulation and realize that outside data is not protected by code. There is no formalized notion of ensuring access to the data is mediated by a body of code. Rather, there is a design point that says if you have access to the raw contents of a message, you should be able to understand it. Inside data is always encapsulated by the service and its application logic.

Consider data and its relationship to its schema. Outside data is immutable, and each data item's schema remains immutable. Note that the schema may be versioned and the new version applied to subsequent similar data items, but that does not change the fact that once a specific immutable item is created, its schema remains immutable. This is in stark contrast to the mechanisms employed by SQL for inside data. SQL's DDL is designed to allow powerful transformations to existing schema while the database is populated.

Finally, let's consider the desirability of performing business intelligence analysis over the data. Experience shows that those analysis folks want to slice and dice anything they can get their hands on. Existing analytics operate largely over inside data, which will certainly continue as fodder for analy-

sis. But there is little doubt about the utility of analyzing outside data as well.

**The dynamic duo of data representations.** Now, let's compare the strengths and weaknesses of these two representations of data, SQL, and JSON:

- ▶ SQL, with its bounded schema, is fantastic for comparing anything with anything (but only within bounds).

- ▶ JSON, with its unbounded schema, supports independent definitions of schema and data. Extensibility is cool too.

**Consider what it takes to perform arbitrary queries.** *SQL is outstanding* because of its value-based nature and tightly controlled schema, which ensure alignment of the values, facilitating the comparison semantics that underlie queries.

*JSON is problematic* because of schema inconsistency. It is precisely the independence of the definition that poses the challenges of alignment of the values. Also, the hierarchical shape and forms of the data may also be a headache. Still, you *can* project consistent schema in a form easily queried. It might be a lossy projection where not all the knowledge is available to be queried.

**Consider independent definition of shared data.** *SQL is impossible* because it has centralized schema. As already discussed, this is intrinsic to its ability to support value-based querying in a tightly controlled environment.

*JSON is outstanding.* It specializes in independent definition of schema and independent generation of documents containing the data. That is a huge strength of JSON and other semi-structured data representations.

**Each model's strength is simultaneously its weakness.** What makes SQL exceptional for querying makes it dreadful for independent definition of shared data. JSON is wonderful for the independent definition, but it stinks for querying (see Figure 6). *You cannot add features to either of these models to address its weaknesses without undermining its strengths.*

## Conclusion

This article describes the impact of services and trust on the treatment of data. It introduces the notions of inside data as distinct from outside data. After discussing the temporal implications of not sharing transactions across the boundaries of services, the article considers the need for immutability and

stability in outside data. This leads to a depiction of outside data as a DAG of data items being independently generated by disparate services.

The article then examines the notion of reference data and its usage patterns in facilitating the interoperation of services. It presents a brief sketch of inside data with a discussion of the challenges of shredding incoming data in the face of extensibility.

Finally, JSON and SQL are seen as representations of data, and their strengths are compared and contrasted. This leads to the conclusion that each of these models has strength in one usage that complements its weakness in another usage. It is common practice today to use JSON to represent data on the outside and SQL to store the data on the inside. Both of these representations are used in a fashion that plays to their respective strengths.

*This is an update to the original paper by the same name presented at the Conference on Innovative Data Systems Research in 2005. At that time, XML was more commonly used than JSON. Similarly, service-oriented architecture (SOA) was used more then, while today, it's more common to say simply, "service." In this article, "service" is used to mean a database encapsulated by its service or application code. It does not mean a microservice. Nomenclature aside, not much has changed.* ■

## Related articles on queue.acm.org

### Beyond Relational Databases

Margo Seltzer

<https://queue.acm.org/detail.cfm?id=1059807>

### The Singular Success of SQL

Pat Helland

<https://queue.acm.org/detail.cfm?id=2983199>

### A co-Relational Model of Data for Large Shared Data Banks

Erik Meijer and Gavin Bierman

<https://queue.acm.org/detail.cfm?id=1961297>

## References

1. Bernstein, P. A., Hadzilacos, V., Goodman, N. *Concurrency Control and Recovery in Database Systems*. Addison-Wesley, 1987; <http://sigmod.org/publications/dblp/db/books/dbtext/bernstein87.html>.
2. Gray, J. and Reuter, A. *Transaction Processing: Concepts and Techniques*. Morgan Kaufmann, 1993.

**Pat Helland** has been implementing transaction systems, databases, application platforms, distributed systems, fault-tolerant systems, and messaging systems since 1978. He currently works at Salesforce.

Copyright held by author/owner.  
Publication rights licensed to ACM.

# ACM Transactions on Computing for Healthcare (HEALTH)

Open for  
Submissions

A multidisciplinary journal for  
high-quality original work on how  
computing is improving healthcare



Computing for Healthcare has emerged as an important and growing research area. By using smart devices, the Internet of Things for health, mobile computing, machine learning, cloud computing and other computing based technologies, computing for healthcare can improve the effectiveness, efficiency, privacy, safety, and security of healthcare (e.g., personalized healthcare, preventive healthcare, ICU without walls, and home hospitals).

*ACM Transactions on Computing for Healthcare* (HEALTH) is the premier journal for the publication of high-quality original research papers, survey papers, and challenge papers that have scientific and technological results pertaining to how computing is improving healthcare. This journal is multidisciplinary, intersecting CS, ECE, mechanical engineering, bio-medical engineering, behavioral and social science, psychology, and the health field, in general. All submissions must show evidence of their contributions to the computing field as informed by healthcare. We do not publish papers on large pilot studies, diseases, or other medical assessments/results that do not have novel computing research results. Datasets and other artifacts needed to support reproducibility of results are highly encouraged. Proposals for special issues are encouraged.

For more  
information  
and to submit  
your work,  
please visit:

[health.acm.org](http://health.acm.org)



Association for  
Computing Machinery



**While millions of students worldwide have enjoyed coding experiences over the last decade, the next challenge is spreading educational values and approaches.**

BY MITCHEL RESNICK AND NATALIE RUSK

## Coding at a Crossroads

THE EDUCATIONAL USE of coding in schools is at a crossroads.

We are at a moment of extraordinary opportunity. A decade ago, our research group wrote an article for *Communications* titled “Scratch: Programming for All.”<sup>15</sup> At the time, our subtitle was aspirational. Now, it is becoming the reality. School systems and policymakers are embracing the idea that coding can and should be for everyone. Countries from Chile to England to South Africa to Japan are introducing coding to all students.

We are also at a moment of extraordinary challenge. In many places, coding is being introduced in ways that undermine its potential and promise. If we do not think carefully about the educational strategies and pedagogies for introducing coding, there is a major risk of disappointment and backlash.

During the past decade, we have seen that it is possible to spread coding experiences to millions of children around the world. But we have also seen that it is much more difficult to spread educational values

and approaches—that is the big challenge for the next decade.

The expansion of coding in education has been catalyzed by new types of programming interfaces (particularly block-based coding<sup>1</sup>), a proliferation of nonprofit initiatives supporting computer-science education (such as Code.org, CSforAll, and Code Club), and a growing array of programmable devices that broaden the range of what students can code (such as micro:bit,<sup>20</sup> robotics kits,<sup>9</sup> and programmable toys<sup>23</sup>).

Our own work on Scratch (Figure 1) has both contributed to and benefitted from this broader trend. When we started developing the Scratch programming language and online community in 2002, our goal was not simply to help children learn to code. We had a broader educational mission. We wanted to provide all children, from all backgrounds, with opportunities to learn to think creatively, reason systematically, and work collaboratively. These skills are essential for everyone in today’s fast-changing world, not just those planning to become engineers and computing professionals. And these same skills are valuable in all aspects of life, not just for success in the workplace but also for personal fulfillment and civic engagement.<sup>13</sup>

The use of Scratch has been growing rapidly throughout the world: in the past year, more than 20 million young people created Scratch projects (Figure 2). Scratch began with use primarily in homes and informal learning settings,<sup>11</sup> but use in schools has expanded to more than half of all Scratch activity. Around the

### » key insights

- In many educational settings, coding is introduced in narrow ways that focus primarily on teaching specific concepts, rather than supporting students in developing the creativity, collaboration, and communication skills needed to thrive in today’s fast-changing world.
- For students to develop computational fluency and creative thinking skills, they need opportunities to create projects, based on their passions, in collaboration with peers, in a playful spirit.

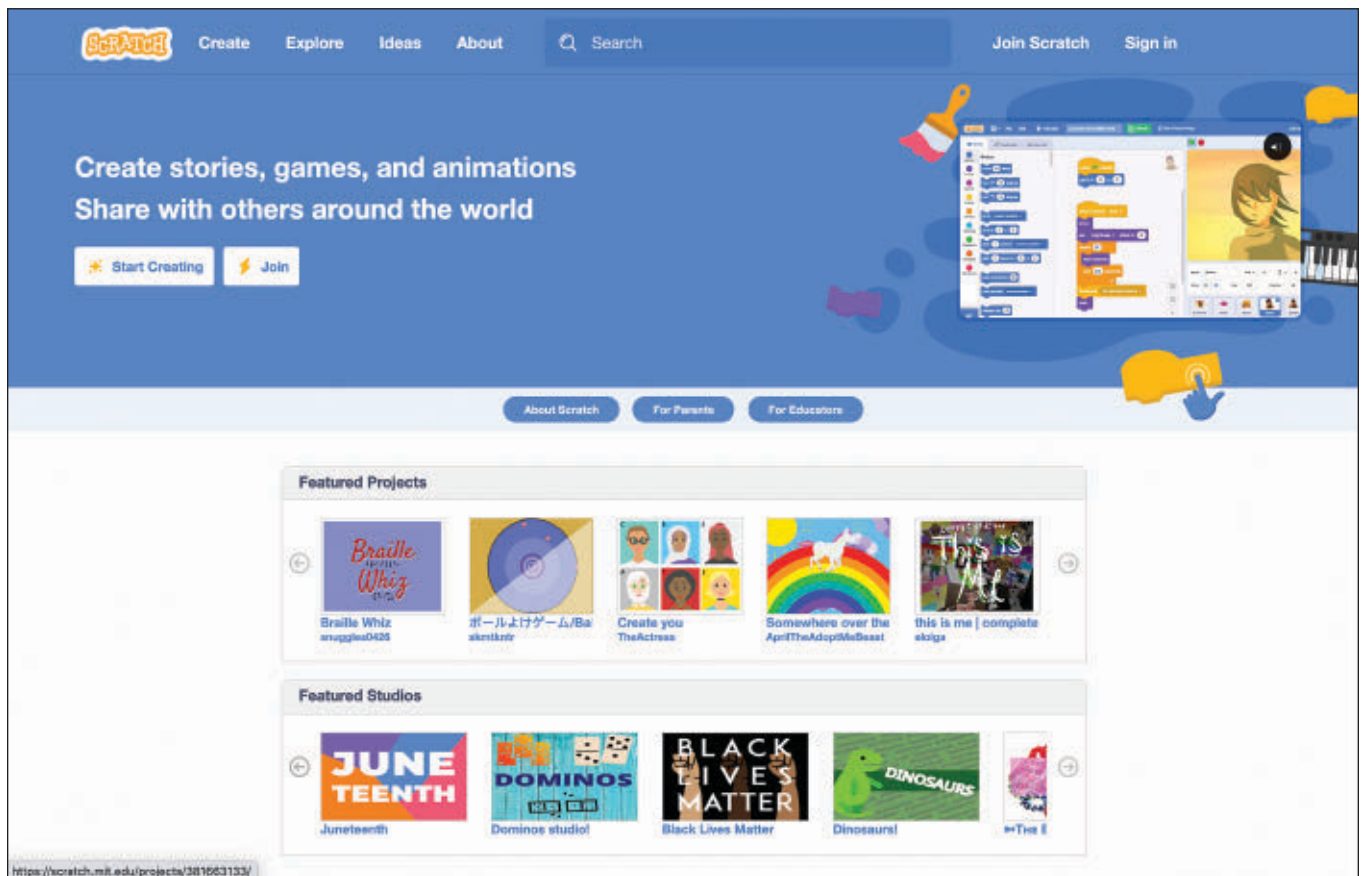


Figure 1. The Scratch website in June 2020.

world, young people are using Scratch in a wide variety of ways. For example:

- ▶ middle-school students across several countries created Scratch projects illustrating their visions for how technological innovations would transform society by the year 2050;

- ▶ thousands of young people created Scratch animations against racism and in support of the Black Lives Matter movement;

- ▶ an elementary-school teacher in Mexico integrated Scratch into a science unit on butterflies, with students creating animations of the butterfly life cycle and robotic models of butterfly motion, based on their observations of real butterflies;

- ▶ students from around the world created a studio called #ProtectOurEarth where they shared hundreds of projects highlighting issues related to climate change, including a game where you guide a polar bear across the melting Arctic ice caps.

### Opportunities and Challenges

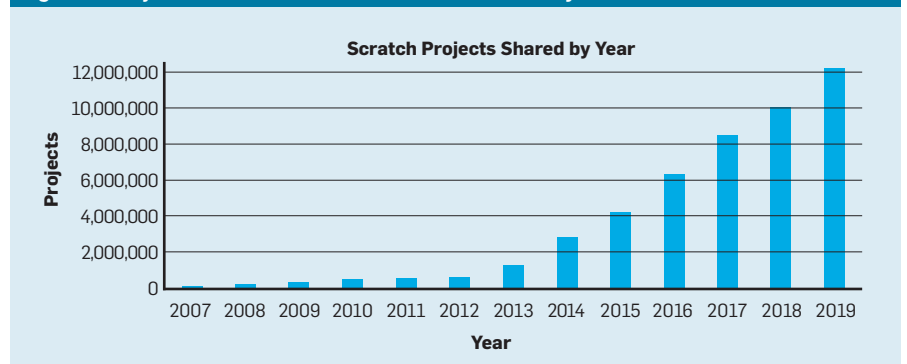
In the process of creating and sharing projects like these, students are not just

learning to code, they are coding to learn. They are not only learning important mathematical and computational concepts, they are also deepening their understanding of ideas in other disciplines and developing a broad range of problem-solving, design, collaboration, and communication skills.<sup>7,16</sup>

Unfortunately, in many educational settings, coding is introduced in much more limited and constrained ways, so that students do not have the opportunity to experience the full conceptual and expressive powers of coding. Here are some of the challenges:

- ▶ Too often, schools are introducing students to computer science by teaching them definitions of words associated with computing, without providing them with opportunities to learn and apply computational concepts and practices in the context of meaningful activities. For example, some school districts introduce computing to elementary-school students by teaching them the definition of the word “algorithm” and the differences between hardware and software, instead of engaging students in active learning through computing activities, such as

Figure 2. Projects shared in the Scratch online community.






coding an animated story or programming a robot to dance.


► Too often, coding is introduced by telling all students to copy the exact same code, rather than encouraging them to experiment, prototype, and debug. On the Scratch website, we once saw 30 identical projects shared at the same time. At first we thought this duplication of projects was a problem with the website, but then we noticed that each project had a different username, and we realized the projects were all from a single classroom, where 30 students had followed the same instructions to make the same project with the same images and same code. Although this classroom activity may have introduced students to the basic mechanics of coding, it did not provide opportunities for creative thinking and problem solving.

► Too often, schools allocate only a brief period of time for learning to code. Within this limited time, students might learn some basic terms and concepts, but they don't have the opportunity to put the ideas to use in a meaningful way, and thus are unlikely to be able to apply the ideas in other contexts and other subjects. And in situations where coding is allocated more time, the curriculum often pushes teachers and students to shift from one coding tool to another, rather than providing time for learning a tool well enough for designing projects, solving problems, and communicating ideas. One large-scale initiative introduced Scratch to fourth-graders for one hour each week, then abruptly shifted to a different coding language. After teachers and students expressed frustration, the curriculum was revised.

► Too often, researchers and educators are adopting automated assessment tools that evaluate student programming projects only by analyzing the code, without considering the project goals, content, design, interface, usability, or documentation. For example, many are using an online Scratch assessment tool that gives students a “computational thinking score” based on the assumption that code with more types of programming blocks is an indication of more advanced computational thinking. This form of assessment doesn't take into



**In our research, we have seen how coding becomes most motivating and meaningful for students when they have opportunities to create their own projects and express their own ideas.**



consideration what the student's program is intended to do, how well it accomplishes the student's goals, whether the code works as intended, whether people are able to interact with it, or how the student's thinking develops over a series of projects. We see greater potential in other research and evaluation approaches, such as those that document and analyze teachers' facilitation practices and students' learning trajectories over time.<sup>6,8</sup>

For coding initiatives to live up to their promise and potential, significant changes are needed in how coding is put into practice in educational systems around the world.

### **Computational Fluency**

In most educational coding initiatives, there is a recognition that the goal should be broader than teaching specific programming techniques. Many educational initiatives are framed around the development of *computational thinking*—that is, helping students learn computer-science concepts and strategies that can be used in solving problems in a wide range of disciplines and contexts.<sup>22</sup>

Computational thinking is certainly a worthy goal, but many initiatives focus too narrowly on teaching concepts out of context or presenting students with problems that have a single correct answer. In our research, we have seen how coding becomes most motivating and meaningful for students when they have opportunities to create their own projects and express their own ideas.<sup>18</sup> Through these experiences, children develop as computational creators as well as computational thinkers. We use the phrase *computational fluency* to describe this ability to use computational technologies to communicate ideas effectively and creatively.

Our ideas about computational fluency have been informed and inspired by the long tradition of educational initiatives and research focused on engaging students in learning to write. Even though most students won't grow up to become professional journalists or novelists, there is a strong consensus that all students should learn to write. Through writing, students develop their ability to organize, express, and share ideas—and they begin to see themselves differently. The Brazilian

educator and activist Paulo Freire led literacy campaigns not simply to help people get jobs, but also to help people learn that “they can make and remake themselves.”<sup>5</sup>

We see the same potential for coding. Most students will not pursue careers as professional programmers or computer scientists but developing fluency with coding is valuable for everyone. As students create their own stories, games, and animations with code, they start to see themselves as creators, developing confidence and pride in their ability to create things and express themselves with new technologies.

Some advocates of computational thinking downplay the value of coding. They argue that there are many other ways to develop computational thinking skills. But we have found that coding can be a particularly effective way for students to become engaged with computational concepts, practices, and perspectives.<sup>2</sup> When students code their own projects, they encounter concepts and problem-solving strategies in a meaningful context, so the knowledge is embedded in a rich web of associations. As a result, students are better able to access and apply the knowledge in new situations.

The Scratch programming language and online community are designed specifically to support the development of computational fluency. Of course, it takes time for students to develop fluency. Many projects in the Scratch online community are very simple or poorly structured, created by students who are just starting to explore the possibilities of coding. But when students have the necessary time and support for developing their fluency, we see how they can grow as both computational thinkers and computational creators.

As an example, we would like to share the story of a Scratch community member named Taryn, who was first introduced to Scratch at her school in South Africa when she was 10 years old. A few years later, in a science class, Taryn used Scratch to program an interactive simulation of the water cycle, including two sliders for controlling the evaporation rates over the sea and over the land. In all, Taryn created a dozen different variables for the project (Figure 3).

Figure 3. Taryn’s Scratch project modeling the water cycle.

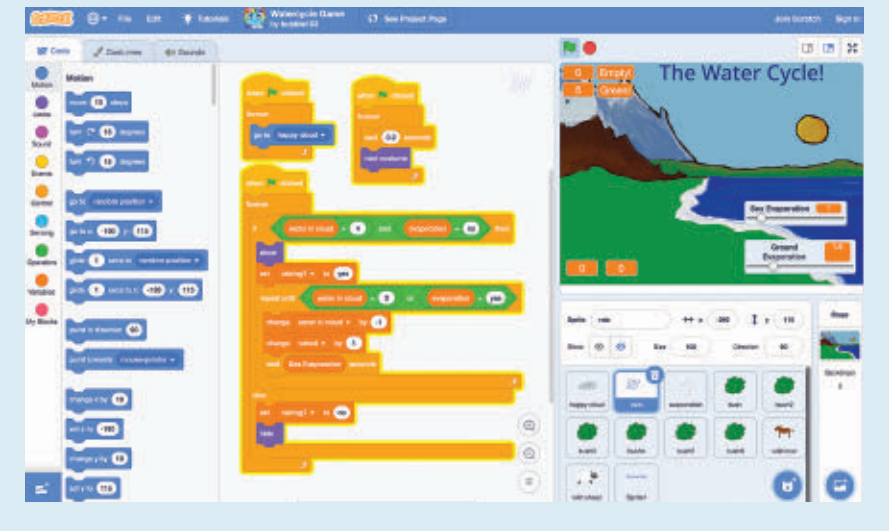


Figure 4. Taryn’s tutorial on how to use variables.

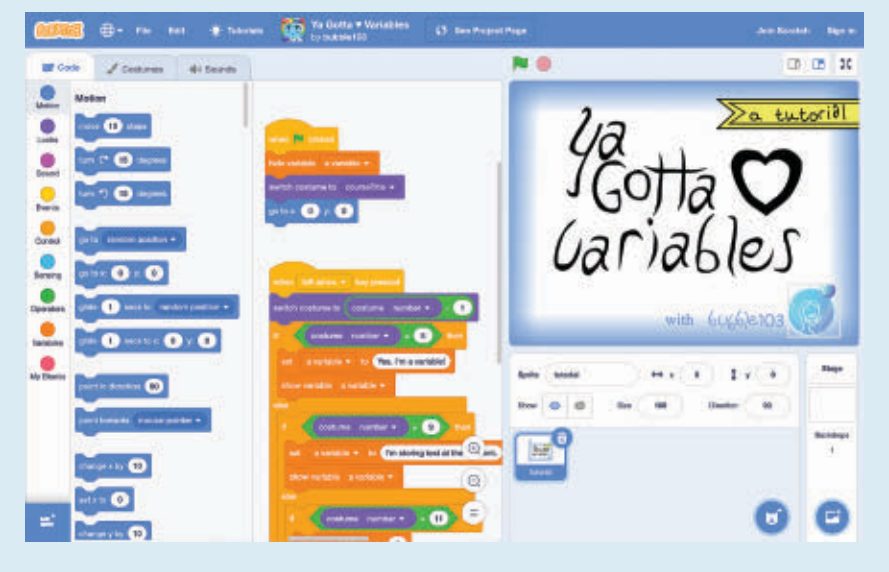


Figure 5. One of Taryn’s Colour Divide animated stories.





Through working on this project, Taryn became inspired to help others learn about variables. She decided to create a tutorial project called *Ya Gotta ♥ Variables* and shared it in the Scratch online community (Figure 4). As she explained in the notes that accompany the project: “I love variables! They’re extremely useful in programming, and I wouldn’t have been able to make most of my projects without them. However, they’re a bit tricky to understand—that’s where this tutorial can help you!” Taryn also encouraged others to experiment: “Have fun playing around and experimenting with variables and booleans! The more you experiment (and fail!), the more you will understand and the easier it will be for you to use variables to make your projects awesome!”

Taryn became well known in the Scratch community through a series of projects called *Colour Divide*, set in a fantasy dystopian world where people are subjected to a test that determines their place in society (Figure 5). Taryn collaborated on the initial *Colour Divide* project with five other students who she met in the online community. For Taryn, the project was a way to explore important social issues. When we interviewed Taryn, she explained: “Growing up, I’ve definitely seen the scars that apartheid has left on my country and the people. I’m really exploring that through the different characters that are a part of this story.”

Taryn described the important role that collaboration played in the development of *Colour Divide*. “I set it up so that other Scratchers could contribute faces and voices and scenery and music. It felt less like something that I was making, more like something that we were making together,” she said. “I’ve just been constantly blown away by the kind of support and collaboration and sharing that happens in the community. That’s one of the main things that keeps me coming back to Scratch every day.”

Through her work on Scratch, Taryn has shifted the way she approaches learning. “I’ve become more confident to try new things and express myself—and more comfortable with taking risks and making mistakes,” she explained. “In other languages, you are

almost too scared to get something wrong and type the wrong thing and be judged. But Scratch it’s like playing, it’s like chucking things together, if they don’t work, that’s fine. And being able to make mistakes is part of the thing that develops creative confidence.”

For us, Taryn’s work serves as an example of how students, through their work on Scratch projects, can develop as both computational creators and computational thinkers. We have seen many other students in the Scratch community go through similar learning trajectories. But many students don’t receive the opportunities or support they need to become fluent with computation and develop as creative thinkers. How can we help more students experience the joys and possibilities of computational fluency?

#### Four Guiding Principles

In our research group, we have developed four guiding principles for supporting creative learning and computational fluency. We call these principles the Four Ps of Creative Learning: *Projects*, *Passion*, *Peers*, and *Play*.<sup>14</sup>

These principles provide a framework to guide the design of technologies, activities, curriculum, communities, and spaces to support coding and learning. Here, we explore the Four Ps of Creative Learning through examples from the Scratch community.

**Projects.** *Provide students with opportunities to work on meaningful projects (not just puzzles or problem-solving activities), so they experience the process of turning an initial idea into a creation that can be shared with others.*

To us, it seems natural to introduce coding to young people in a project-oriented way, so that they learn to express themselves creatively as they learn to code. But many introductions to coding take a very different approach, presenting students with a series of logic puzzles in which they need to program animated characters to move from one location to another. When students successfully solve one puzzle, they can move on to the next. Students undoubtedly learn some useful computational concepts while working on these puzzles. But learning to code by solving logic puzzles is somewhat like learning to write by solving crossword puzzles. That’s not the way to become truly fluent. Just as

students develop fluency with language by writing their own stories (not just playing word games), students develop fluency with coding by creating projects (not just solving puzzles).

Increasingly, schools are shifting to a project-based approach to coding. In one school, for example, fourth-grade students created Scratch projects about the book *Charlotte’s Web*, rather than writing traditional book reports. In one of the projects, a student programmed a pig to move within the scene. To make the pig look further away, the student programmed it to become smaller, applying the art concept of perspective and using mathematical calculations to adjust the size of the pig. The project cut across the curriculum, integrating ideas from language, art, math, and computer science. In other schools, students have designed projects in many different subject areas—creating games about ancient Egypt in history class, modeling DNA replication in biology, and creating animations of haiku poems in language arts.

For teachers, it might be easier to introduce coding through puzzles that tell students whether they have correctly solved the problem or where they went wrong. Managing a project-based classroom can be more challenging, since different students will create different types of projects. Yet it is precisely this opportunity for developing an idea from initial conception to shareable project that enables young people to develop as creative thinkers and problem solvers.<sup>14</sup>

**Passion.** *Allow students to work on projects connected to their interests. They will work longer and harder—and learn more in the process.*

We designed Scratch to support a wide range of projects and interests—from art, music, and animations, to games, stories, and simulations. We also made sure students can customize and personalize their projects, by bringing in their own images and sounds.

Why is this important? Different children have different interests, come from different cultures, and think in different styles. Supporting diverse pathways into Scratch is important to ensure that all children, from all backgrounds, can work on Scratch projects that are relevant and meaningful to them. On the Scratch website, you can see a wide

diversity of projects, everything from interactive newsletters to dance tutorials to historical dress-up games to musical beat machines. That's an indication that Scratch is supporting students with a wide range of different interests and passions. Similarly, when evaluating Scratch classes or workshops, we use diversity of projects as a measure of success—an indication that children are working on projects they care about.

In an influential paper from the 1990s, Sherry Turkle and Seymour Papert emphasized that encouraging diverse styles of thinking and programming is essential for promoting equity and developing a more inclusive computer culture.<sup>21</sup> As they wrote:

“The computer is an expressive medium that different people can make their own in their own way ... The diversity of approaches to programming suggests that equal access to even the most basic elements of computation requires accepting the validity of multiple ways of knowing and thinking, an epistemological pluralism.”

We often refer to this idea with the phrase “many paths, many styles.” Some students make elaborate plans, others explore and tinker. Some students enjoy telling stories, others enjoy making patterns. Some students are excited about animals, others are excited about sports. To ensure coding is for all, it is important to support these diverse entry points and approaches.


**Peers.** *Encourage collaboration and sharing, and help students learn to build on the work of others.*

When our research group launched the Scratch programming language in 2007, we launched the Scratch online community at the same time. We wanted to support the social side of learning, providing students with opportunities to learn with and from one another. The online community has grown into a dynamic space where young people collaborate with one another, sharing more than one million projects and posting more than three million comments each month.

We have learned from Scratchers just how important the online community is for motivating their ongoing participation.<sup>18</sup> As one Scratcher explained: “I would've quit earlier, but then I made friends ... Of course, I had friends in real life, but having friends in other



**The online community has grown into a dynamic space where young people collaborate with one another, sharing more than one million projects and posting more than three million comments each month.**



countries with the same interests kept me coming back to talk to them.”

Young people talk about multiple reasons why the Scratch online community matters to them:

- ▶ The community provides *audience*: When young people share projects they have made, they get feedback, encouragement, and suggestions from peers in the community.

- ▶ The community provides *inspiration*: By looking at other projects on the website, young people get new ideas for their own projects.

- ▶ The community provides *connection*: Young people make friends and meet others with shared interests from other cities and countries.

As a young person in the online community reflected:

“When I used the website, I got interested in the projects of others. This is largely how I learned Scratch: through remixing and sharing and creating. I made many friends here, who remix my projects, give comments, and have taught me new things.”

As participation in the Scratch community has grown, young people have collaborated in ways beyond what we had originally anticipated. More and more young people have taken the initiative to connect, coordinate, and collaborate on projects and activities. About a quarter of all projects on the Scratch website are remixes, in which students modify or add code to existing projects.<sup>4</sup> Some students form collaborative groups to create complex games and animations that none could have created on their own. Other students have learned how to create projects through crowdsourcing, asking others in the community to contribute code, images, or sound clips.<sup>17</sup>

A few years ago, a college physics professor told us his children had become actively involved in the Scratch community. We expected he would go on to tell us about the coding skills and computational ideas they were learning. But that's not what interested him most. Rather, he was excited that his children were participating in an open knowledge-building community. “It's like the scientific community,” he explained. “Kids are constantly sharing ideas and building on one another's work. They're learning how the scientific community works.”




**Play.** *Create an environment where students feel safe to take risks, try new things, and experiment playfully.*


Scratch is designed to encourage playful experimentation and tinkering. As with LEGO bricks, it is easy to snap together Scratch programming blocks to try out new ideas, and it is also easy to take them apart to revise and iterate. Just click on a stack of Scratch blocks, and the code runs immediately. There are no error messages in the Scratch programming editor. Instead, many children learn new coding strategies by playfully experimenting with different combinations of Scratch blocks, seeing what happens when their code runs, iteratively revising their code, and looking at code in other projects. We view “play” not as an activity but as an attitude: a willingness to experiment, take risks, and try new things.

When we have interviewed long-time Scratchers, we have found that many became engaged in coding by “messing around” with Scratch.<sup>16</sup> For example, a long-time Scratcher explained that he learned about variables, events, and other coding concepts “just by experimenting.” Although it might seem more efficient to teach concepts through direct instruction, we have seen that many students become more engaged and gain a greater sense of agency and confidence when they learn through playful experimentation and exploration. We do offer tutorials on the Scratch website, but the tutorials are designed to encourage students to incorporate their own ideas and make their own variations, not just follow step-by-step instructions.

The Scratch community guidelines emphasize the importance of being respectful and friendly, and clearly state that Scratch “welcomes people of all ages, races, ethnicities, religions, abilities, sexual orientations, and gender identities.”<sup>19</sup> Respectful communication and inclusiveness have become norms that experienced participants communicate to newcomers and others.<sup>10</sup> A respectful community is essential for accomplishing our goals with Scratch. When people feel they are surrounded by caring, respectful peers, they are much more likely to play—that is, to try new things and take the risks that are an essential part of the creative process.



**We have been encouraged to see a growing number of teachers and schools are finding ways to integrate creative, expressive approaches to coding into their classroom practices.**



### Putting the Four Ps into Practice

From our observations of Scratch activities around the world over the past decade, we have seen the value of Projects, Passion, Peers, and Play in supporting the development of computational fluency. But we have also seen that it is not easy to put these four principles into practice within the realities of today’s standards-based, assessment-driven classrooms.

We have been encouraged to see a growing number of teachers and schools are finding ways to integrate creative, expressive approaches to coding into their classroom practices. In a public high school in Tacoma, WA, for example, computer-science teacher Jaleesa Trapp wanted to provide her students with an opportunity to learn computational concepts in the context of projects that would be meaningful to them. Jaleesa noticed that many of her students enjoyed watching how-to videos online, so she proposed that they use Scratch to create their own how-to tutorials.

The students created a wide range of projects: how to crochet, how to use a 3D printer, and how to make a video game, among others. The students designed their projects to make them accessible to users with diverse abilities. To create their projects, students needed to research their topics, develop prototype tutorials, test out their prototypes with other students, revise their projects, and finally present their projects to friends and family, as well as sharing with a broader audience online.

This activity was well-aligned with the four Ps, since students were working on projects based on their passions, in collaboration with peers, in a playful spirit. But the activity was also well-aligned with computer science and engineering standards, since it involved iterative design, testing, debugging, and refinement of computer programs.<sup>3,12</sup> Students gained an understanding of important computational concepts and practices (such as using control structures and improving usability) through working on their projects.

Jaleesa also wanted an assessment method that would be meaningful to the students. So, before they started designing, she asked the students to help develop a rubric for evaluating their projects. They began by identifying the features of how-to videos that

they valued and decided together which criteria were most important to include in the rubric. By contributing to the criteria for assessment, the students developed a shared understanding of the goals, and they were invested in meeting them.

Jaleesa noted that many computer-science initiatives evaluate students based on how many different programming blocks they use in their projects. Jaleesa worried that focusing on this metric might lead students to simply add programming blocks to fulfill a requirement, without understanding the purpose of the different blocks. Instead, the students in Jaleesa’s class used a wide variety of programming blocks in an authentic way. Because students were designing how-to projects to support accessibility, they naturally needed to coordinate multiple events, incorporate multiple types of media, and respond to different types of user input.

### The Next Decade

We are at a moment of great opportunity but also great challenge. Even as new technologies have flowed into schools and as new coding initiatives have been adopted, the core structures of most educational institutions have remained largely unchanged. If new technologies and new coding initiatives are to live up to their promise, we must break down structural barriers in the educational system.

We need to break down barriers across disciplines, providing students with opportunities to work on projects that integrate science, art, engineering, and design. We need to break down barriers across age, allowing people of all ages to learn with and from one another. We need to break down barriers across space, connecting activities in schools, community centers, and homes. And we need to break down barriers across time, enabling children to work on interest-based projects for weeks or months, rather than squeezing projects into the constraints of a class period or curriculum unit.

Breaking down these structural barriers is difficult. It requires a shift in the ways people think about education and learning. People need to view education not as a way to deliver information, but rather as a way to support stu-

dents in exploring, experimenting, and expressing themselves, so that students can develop the creativity, collaboration, and communication skills that are needed to thrive in today’s fast-changing world.

These changes in structures and mindsets will require efforts by many people, in many places, at many levels. There are already teachers, schools, and even entire districts that are implementing new, creative approaches to coding and learning. We need to build on these examples to support broader change. No individual policy or individual school or individual technology can bring about change on its own. We need a movement in which people in all parts of the educational ecosystem—educators, administrators, researchers, curriculum developers, toolmakers, and policymakers—think about coding in new ways and think about learning in new ways.

We are at a crossroads. Ten years from now, we hope we can look back and report on a decade of educational change, in which schools have provided students with the time, space, support, and encouragement they need to become fluent with new technologies, so that they can help shape tomorrow’s society.

### Acknowledgments

Many people have contributed to the design, development, and support of Scratch, particularly members of the Lifelong Kindergarten Group at the MIT Media Lab and the Scratch Team at the Scratch Foundation. We are grateful to the National Science Foundation for supporting the initial research and development of Scratch, and to the Siegel Family Endowment, LEGO Foundation, and other supporters for making it possible to make Scratch available for free for young people and educators around the world. C

### References

1. Bau, D., Gray, J., Kelleher, C., Sheldon, J. and Turbak, F. Learnable programming: blocks and beyond. *Commun. ACM* 60, 6 (Jun. 2017), 72–80; <https://dl.acm.org/citation.cfm?doid=3098997.3015455>
2. Brennan, K. and Resnick, M. Using artifact-based interviews to study the development of computational thinking in interactive media design. *Annual Meeting of the American Educational Research Association*, Vancouver, B.C. 2012.
3. Computer Science Teachers Association. CSTA K-12 Computer Science Standards, 2017; <http://www.csteachers.org/standards>
4. Dasgupta, W.H., Monroy-Hernández, A. and Hill, B.M. Remixing as a pathway to computational thinking. In *Proceedings of the 19th ACM Conference on Computer-*

*Supported Cooperative Work & Social Computing* (2016). ACM, New York, 1438–1449. <https://doi.org/10.1145/2818048.2819984>

5. Freire, P. *Pedagogy of Indignation*. Paradigm, Boulder, CO, 2014.
6. Israel, M., Pearson, J.N., Tapia, T., Wherfel, Q.M., and Reese, G. Supporting all learners in school-wide computational thinking: A cross-case qualitative analysis. *Computers & Education*, 82 (Mar. 2015), 263–279; <https://doi.org/10.1016/j.compedu.2014.11.022>
7. Kafai, Y.B. and Burke, Q. *Connected Code: Why Children Need to Learn Programming*. MIT Press, Cambridge, MA, 2014.
8. Ke, F. An implementation of design-based learning through creating educational computer games: A case study on mathematics learning during design and computing. *Computers & Education*, 73 (Apr. 2014), 26–39.
9. Khine, M.S. *Robotics in STEM Education*. Springer, 2017; <https://doi.org/10.1007/978-3-319-57786-9>
10. Lombana-Bermudez, A. Moderation and sense of community in a youth-oriented online platform. 2017; <https://bit.ly/2NfpxEl>
11. Maloney, J., Peppler, K., Kafai, Y., Resnick, M., and Rusk, N. Programming by choice: Urban youth learning programming with Scratch. *ACM SIGCSE Bulletin* 40, 1 (Mar. 2008), 367–371.
12. NGSS Lead States. *Next Generation Science Standards: For States, by States*. National Academies Press, Washington, D.C., 2013.
13. National Research Council. *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. National Academies Press, Washington, D.C., 2013
14. Resnick, M. *Lifelong Kindergarten: Cultivating Creativity through Projects, Passion, Peers, and Play*. MIT Press, Cambridge, MA, 2017.
15. Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., Millner, A., Rosenbaum, E., Silver, J., Silverman, B., and Kafai, Y. Scratch: Programming for all. *Commun. ACM* 52, 11 (Nov. 2009), 60–67.
16. Roque, R. and Rusk, N. Youth perspectives on their development in a coding community. *Info. Learning Sci.* (Apr. 2019); <https://doi.org/10.1108/ILS-05-2018-0038>
17. Roque, R., Rusk, N., and Resnick, M. 2016. Supporting diverse and creative collaboration in the Scratch online community. *Mass Collaboration and Education*. U. Cress, H. Jeong, and J. Maskaliuk (Eds.) Springer, Cham, Switzerland. 241–256; [https://doi.org/10.1007/978-3-319-13536-6\\_12](https://doi.org/10.1007/978-3-319-13536-6_12)
18. Rusk, N. Motivation for making. *Makeology: Makers as Learners*, K. Peppler, E. Rosenfeld Halverson, and Y.B. Kafai (Eds.). Routledge, New York, NY, 85–108.
19. Scratch Community Guidelines. 2018; [http://scratch.mit.edu/community\\_guidelines/](http://scratch.mit.edu/community_guidelines/)
20. Sentance, S., Waite, J., Hodges, S., MacLeod, E., and Yeomans, L. ‘Creating cool stuff’: Pupils’ experience of the BBC micro:bit. In *Proceedings of the 2017 ACM SIGCSE* (Seattle, WA) 531–536.
21. Turkle, S. and Papert, S. Epistemological pluralism: Styles and voices within the computer culture. *SIGNS*: 16, 1 (1990), 128–157.
22. Wing, J.M. Computational thinking. *Commun. ACM* 49, 3 (Mar. 2006), 33–35.
23. Yu, J. and Roque, R. A review of computational toys and kits for young children. *Int’l J. Child-Computer Interaction*. (Jul. 2019); <https://doi.org/10.1016/j.jcci.2019.04.001>

**Mitchel Resnick** (mres@media.mit.edu) is Professor of Learning Research at the MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA.

**Natalie Rusk** (nrusk@media.mit.edu) is Research Scientist in the Lifelong Kindergarten group at the MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA.

Copyright held by author/owner.  
Publication rights licensed to ACM.



Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/coding-at-a-crossroads>



## Exploring the theoretical and practical aspects of the graph isomorphism problem.

BY MARTIN GROHE AND PASCAL SCHWEITZER

# The Graph Isomorphism Problem

DECIDING WHETHER TWO graphs are structurally identical, or *isomorphic*, is a classical algorithmic problem that has been studied since the early days of computing. Applications span a broad field of areas ranging from chemistry (Figure 1) to computer vision. Closely related is the problem of detecting symmetries of graphs and of general combinatorial structures. Again this has many application domains, for example, combinatorial optimization, the generation of combinatorial structures, and the computation of normal forms. On the more theoretical side, the problem is of central interest in areas such as logic, algorithmic group theory, and quantum computing.

Graph isomorphism (GI) gained prominence in the theory community in the 1970s, when it emerged as one of the few natural problems in the complexity class NP that could neither be classified as being hard (NP-complete) nor shown to be solvable with an efficient algorithm (that is, a polynomial-time

algorithm). It was mentioned numerous times as an open problem, in particular already in Karp's seminal 1972 paper<sup>23</sup> on NP-completeness as well as in Garey and Johnson's influential book on computers and intractability.<sup>15</sup> Since then, determining the precise computational complexity of GI has been regarded a major open problem in theoretical computer science.

In a recent breakthrough,<sup>3</sup> Babai proved that GI is solvable in *quasipolynomial time*. This means that on  $n$ -vertex input graphs, Babai's algorithm runs in time  $n^{p(\log n)}$  for some polynomial  $p(X)$ . This can be interpreted as the problem being almost efficiently solvable—theoretically.

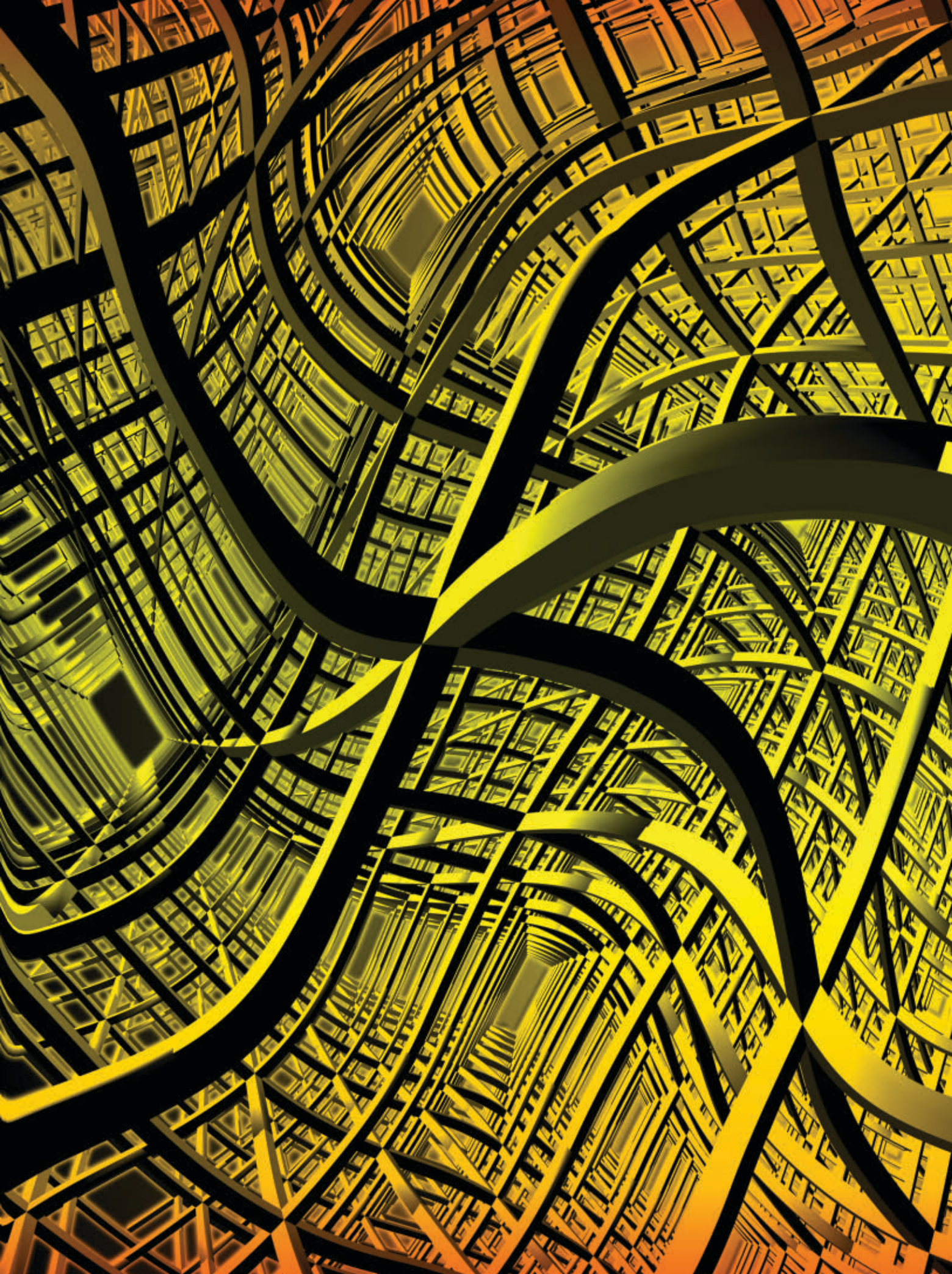
In this paper, we will survey both theoretical and practical aspects of the graph isomorphism problem, paying particular attention to the developments that led to Babai's result.

**Historical development.** Graph isomorphism as a computational problem first appears in the chemical documentation literature of the 1950s (for example, Ray and Kirsch<sup>35</sup>) as the problem of matching a molecular graph (see Figure 1) against a database of such graphs. The earliest computer science reference we are aware of is due to Unger,<sup>39</sup> incidentally also in the *Communications of the ACM*. Maybe the first important step on the theoretical side was Hopcroft and Tarjan's  $O(n \log n)$  isomorphism algorithm for planar graphs.<sup>22</sup> As the question whether GI is NP-complete gained prominence, it was realized that GI has aspects that distinguish it from most NP-complete

### » key insights

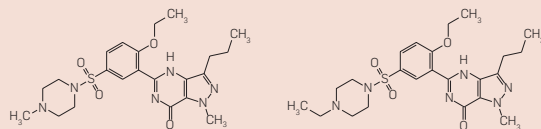
- With its many practical and theoretical applications the graph isomorphism problem remains one of the most important unresolved problems in theoretical computer science.
- A recent breakthrough by László Babai shows that the problem is almost efficiently solvable—theoretically.
- We are only starting to explore the potential of the wealth of new ideas the recent advances bring to the field.



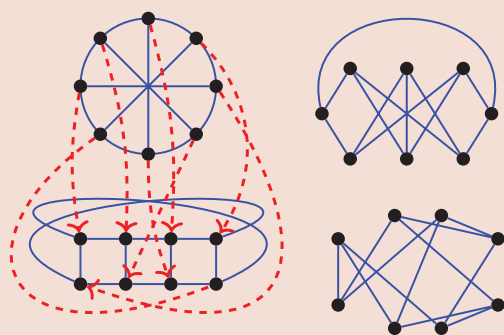




**Figure 1. Nonisomorphic molecular graphs.**



**Figure 2. Four isomorphic graphs. The red arrows indicate an isomorphism between the first and the third graph.**



problems. In particular, counting the number of isomorphisms between two graphs is not harder than deciding if there is an isomorphism (see Mathon<sup>29</sup>). Babai et al.<sup>8</sup> showed that GI is easy on average with respect to a uniform distribution of input graphs. In fact, this can be extended to most other random graph distributions.

A first wave of substantial progress came in 1979–1980 with Babai<sup>2</sup> and Luks’s<sup>26</sup> introduction of group theoretic techniques. In his paper, Luks<sup>26</sup> showed that isomorphism can be decided in polynomial time on graph classes of bounded degree (that is, the number of edges incident with each vertex is bounded), and Luks laid the foundation for much of the subsequent work on graph isomorphism algorithms by introducing a general divide-and-conquer framework. (We will discuss this framework in some detail.) A combination of Luks’s group theoretic framework with a clever combinatorial trick by Zemlyachenko led to a moderately exponential algorithm for graph isomorphism (see Babai and Luks<sup>10</sup>). The best bound was  $2^{O(\sqrt{n \log n})}$ , established by Luks in 1983 (see Babai et al.<sup>9</sup>). This remained the best known bound until Babai’s recent breakthrough.

Around the same time, McKay developed his isomorphism tool Nauty,<sup>30</sup>

which marks a breakthrough in practical isomorphism testing.

In the mid-1980s, another fascinating facet of the complexity of GI was discovered. Using the newly developed machinery of interactive proof systems, it was shown that the complement of GI has short zero-knowledge proofs<sup>16</sup> and this was seen as another indication that GI is not NP-complete. (If GI is NP-complete, then the so-called polynomial hierarchy of complexity classes above NP collapses to its second level, which is regarded as unlikely.) On the other hand, Torán<sup>38</sup> proved that GI is hard to solve when the available memory is quite limited (specifically it is hard for nondeterministic logarithmic space under logspace reductions), which gives us at least some complexity theoretic lower bound.

As the group theoretic methods have been introduced in the early 1980s, they have been continually refined. The underlying group theory has progressed (for example, Babai et al.<sup>5,9</sup>), the complexity of the group theoretic problems has been analyzed in detail (for example, Luks<sup>27</sup> and Seress<sup>37</sup>), and the scope of the methods has been expanded to other structures (for example, Babai et al.<sup>7</sup>). Another active strand of research has been the design of efficient algorithms for GI restricted to graphs with specific properties (for example, Babai et al.<sup>6</sup>, Grohe and Marx,<sup>19</sup> Lokshtanov et al.,<sup>25</sup>

Ponomarenko<sup>34</sup>). More recently, there has also been work on memory restricted algorithms (for example, Datta et al.<sup>14</sup>).

But no real progress on the general isomorphism problem was made until—out of the blue—Babai published his quasipolynomial-time algorithm in 2015.

After this historical overview, let us get slightly more concrete.

**Isomorphisms, automorphisms, and canonical forms.**

An *isomorphism* from a graph  $G = (V, E)$  to a graph  $H = (W, F)$  is a one-to-one mapping  $\pi$  from the vertices of the first graph  $V$  onto the vertices of the second graph  $W$  that preserves adjacency and nonadjacency, that is,  $uv \in E$  if and only if  $\pi(u)\pi(v) \in F$  for all pairs  $uv$  of vertices in  $V$  (Figure 2).

An *automorphism*, or a symmetry, of a graph  $G$  is an isomorphism from  $G$  to  $G$  itself. For example, all  $n!$  permutations of the vertex set of a complete graph  $K_n$  on  $n$  vertices are automorphisms. By comparison, an (undirected) path of length  $n$  only has two automorphisms, the trivial identity mapping, and the mapping that flips the ends of the path. The collection of all automorphisms of  $G$  forms a mathematical structure known as a (permutation) group. As the example of the complete graph shows, automorphism groups can get very large, exponentially large in the number of vertices, but fortunately every permutation group has a generating set linear in the size of the permutation domain (that is, the set of objects being permuted). This allows us to work with automorphism groups efficiently as long as they are represented by sufficiently small generating sets. The problem GI of deciding whether two graphs are isomorphic and the problem of computing a generating set for the automorphism group of a graph (AUT) have the same computational complexity, or more precisely, can be reduced to each other by polynomial-time reductions (see Mathon<sup>29</sup>).

Another important related problem is the graph canonization problem. A *canonical form*  $\gamma$  maps each graph  $G$  to an isomorphic graph  $\gamma(G)$  in such a way that if graphs  $G$  and  $H$  are isomorphic then the graphs  $\gamma(G)$  and  $\gamma(H)$  are identical (not just isomorphic).

**Figure 3. The color refinement algorithm.**

**Color Refinement**

**Input:** Graph  $G$

**Initialization:** All vertices get the same color.

**Refinement Step:** For all colors  $c$  in the current coloring and all nodes  $v, w$  of color  $c$ , nodes  $v$  and  $w$  get different colors in the new coloring if there is some color  $d$  such that  $v$  and  $w$  have different numbers of neighbors of color  $d$ .

The refinement is repeated until the coloring is stable, then the stable coloring is returned.

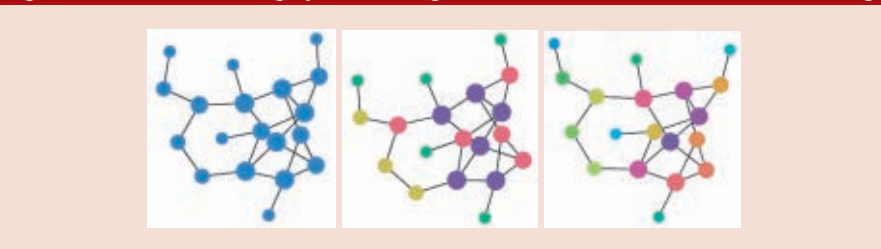
Observe that a canonical form  $\gamma$  yields an isomorphism test: given  $G, H$ , compute  $\gamma(G)$  and  $\gamma(H)$  and check if they are identical. In practical applications, canonical forms are often preferable over isomorphism tests. It is an open problem whether these two problems are actually equivalent (for example, whether the existence of a polynomial-time isomorphism algorithm would yield the existence of a polynomial-time computable canonical form). However, typically for graph classes for which we know a polynomial-time isomorphism algorithm, we also have a polynomial-time canonization algorithm. Sometimes, the extension from isomorphism testing to canonization is straightforward; sometimes it requires extra work (for example, Babai,<sup>4</sup> Babai and Luks,<sup>10</sup> Schweitzer and Wiebking<sup>36</sup>).

**Combinatorial Algorithms**

To establish that two graphs are isomorphic, we can try to find an isomorphism. To establish that the graphs are nonisomorphic, we can try to find a “certificate” of nonisomorphism. For example, we can count vertices, edges, and triangles in both graphs; if any of these counts differ, the graphs are nonisomorphic. Or we can look at the degrees of the vertices. If there is some  $d$  such that the two graphs have a different number of vertices of degree  $d$ , the graphs are nonisomorphic.

The *Weisfeiler-Leman algorithm* provides a systematic approach to generate such certificates of nonisomorphism in an efficient way. Actually, it is a whole family of algorithms, parameterized by a positive integer, the *dimension*.

**Figure 4. Color refinement: a graph, its coloring after 1 refinement round, and the final coloring.**



**Color refinement.** We start by describing the 1-dimensional version, which is commonly known as *color refinement* or *naive vertex classification*. It is one of the most basic ideas in graph isomorphism testing that has been reinvented several times; the oldest published version that we are aware of can be found in Morgan.<sup>32</sup> Color refinement is an important subroutine of almost all practical graph isomorphism tools, and it is also a building block for many theoretical results.

The color refinement algorithm, displayed in Figure 3, iteratively computes a coloring of the vertices of a graph. The actual colors used are irrelevant, what matters is the partition of the vertices into color classes. The final coloring has the property that any two vertices of the same color have the same number of neighbors in each color class. Figure 4 shows an example.

The coloring computed by the algorithm is *isomorphism invariant*, which means that if we run it on two isomorphic graphs, the resulting colored graphs will still be isomorphic and in particular have the same numbers of nodes of each color. Thus, if we run the algorithm on two graphs and find that they have distinct numbers of vertices of some color, we have produced a certificate of nonisomorphism. If this is the case, we say that color refinement *distinguishes* the two graphs.

Unfortunately, color refinement does not distinguish all nonisomorphic graphs. Figure 5 shows a simple example. But, remarkably, color refinement does distinguish *almost all* graphs, in a precise probabilistic sense.<sup>8</sup> This, together with its efficiency, is what makes color refinement so useful as a subroutine of practical isomorphism tools.

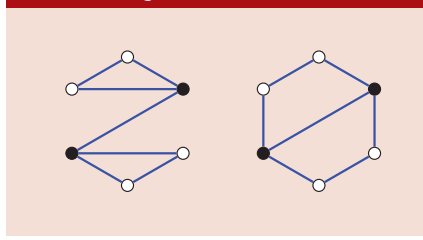
The reader may have noticed that color refinement is very similar to other partitioning algorithms, in particular the standard algorithm for minimizing

deterministic finite automata. Borrowing ideas from Hopcroft’s DFA minimization algorithm,<sup>21</sup> color refinement can be implemented to run in time  $O((n+m)\log n)$ , where  $n$  is the number of vertices and  $m$  is the number of edges of the input graph.<sup>13</sup> Thus, color refinement is indeed very efficient.

**Weisfeiler-Leman.** We have seen that color refinement is not a complete isomorphism test: it fails to distinguish extremely simple nonisomorphic graphs such as those shown in Figure 5. The *k-dimensional Weisfeiler-Leman algorithm* (*k-WL*) is based on the same iterative-refinement idea as color refinement, but is significantly more powerful. Instead of vertices, *k-WL* colors *k*-tuples of vertices of a graph. Initially, each *k*-tuple is “colored” by the isomorphism type of the subgraph it induces. Then in the refinement rounds, the color information is propagated between “adjacent” tuples that only differ in one coordinate (details can be found in Cai et al.<sup>11</sup>). The 2-dimensional version of the algorithm is due to Weisfeiler and Leman;<sup>40</sup> the generalization to higher dimensions is due to Faradzev, Zemlyachenko, Babai, and Mathon (see Cai et al.<sup>11</sup>). If implemented using similar ideas as for color refinement, *k-WL* runs in time  $O(n^{k+1} \log n)$ .

Higher-dimensional WL is very powerful. Indeed, it is highly nontrivial to find nonisomorphic graphs that are not distinguished by 3-WL. It took a

**Figure 5. Two nonisomorphic graphs that are not distinguished by color refinement. Color refinement computes the black-white coloring of the vertices.**





now seminal paper, by Cai et al.<sup>11</sup>, to prove that for every  $k$ , there are nonisomorphic graphs  $G_k, H_k$  that are not distinguished by  $k$ -WL. Indeed, these graphs, known as the *CFI graphs*, have size  $O(k)$  and are 3-regular.

It turns out that many natural graph classes do not admit the CFI-graph construction and a low-dimensional WL is a complete isomorphism test. In particular, for all graph classes  $\mathcal{C}$  that exclude some fixed graph as a minor, there is a constant  $k$  such that  $k$ -WL distinguishes all nonisomorphic graphs in  $\mathcal{C}$ .<sup>17</sup> This includes the class of planar graphs, for which 3-WL suffices.<sup>24</sup>

The Weisfeiler-Leman algorithm is remarkably robust. It not only subsumes most combinatorial ideas for graph isomorphism testing but also has a natural characterization in terms of logic.<sup>11</sup> Surprisingly, it also corresponds to a natural isomorphism test based on linear programming<sup>1</sup> and subsumes various approaches to GI

based on algebraic and mathematical programming techniques.

**Group theoretic algorithms**

Although most isomorphism algorithms devised over the years are subsumed by the Weisfeiler-Leman algorithm, this is not the case for the group theoretic approach.<sup>2, 11</sup> The first application of algorithmic group theory to isomorphism testing was given by Babai.<sup>2</sup> Subsequently, Luks<sup>26</sup> used a group theoretic approach to devise a polynomial-time isomorphism test for graphs of bounded degree.

As GI and the automorphism group problem AUT are polynomially equivalent (see Mathon<sup>29</sup>), it suffices to solve the latter. Starting with a suitable group of permutations, we want to compute within it the automorphism group  $A$  of interest (technically we want to compute a certain set-stabilizer or the solution to a string isomorphism problem, on which we will not elaborate here). We continually maintain an enclosing group  $\Gamma \geq A$  containing all automorphisms as a subgroup. Our strategy is to iteratively shrink  $\Gamma$  until it agrees with  $A$ .

To shrink the group  $\Gamma$ , in case the permutation group  $\Gamma$  has more than one orbit (see Figure 6), we process orbits sequentially.

If the group has only one orbit, we exploit so-called blocks whenever they exist. A *block* of a permutation group  $\Gamma \leq \text{Sym}(\Omega)$  is a subset always mapped to itself or somewhere else entirely, that is, a set  $B \subseteq \Omega$  of the permutation domain  $\Omega$  such that for all  $\gamma \in \Gamma$  we have  $\gamma(B) = B$  or  $\gamma(B) \cap B = \{\}$ . The set of images of the

block  $\{\gamma(B) | \gamma \in \Gamma\}$  partitions the domain  $\Omega$  into blocks of equal size, which together form a so-called *block system*. The group  $\Gamma$  permutes the blocks of the system and we can consider the induced permutation group  $\Gamma'$  on the blocks. By choosing  $B \subseteq \Omega$  inclusion-wise maximal among the blocks, we can ensure that  $\Gamma'$  does not have any (nontrivial) blocks itself. A group with this property is called *primitive*. Luks argues that in polynomial time, we can reduce the computation of the automorphism group  $A$  to  $|\Gamma'|$  computations, each involving subproblems with significantly smaller orbits, which can then be processed sequentially as mentioned above. In case we started with a primitive group, we use a brute force algorithm, inspecting all permutations in  $\Gamma$  separately.

A crucial observation is now that for graphs of bounded degree, there is a method to guarantee that  $|\Gamma'|$  cannot be too large. Originally Luks presented a more involved argument but a subsequent result by Babai et al.<sup>2</sup> directly shows that  $|\Gamma'|$  is polynomially bounded in the permutation domain size. Overall, this bound implies that the entire procedure runs in polynomial time on graphs of bounded degree.

For general graphs, the bottleneck of this procedure occurs when  $\Gamma'$  is large. In that case,  $\Gamma'$  is a large primitive group. Such groups are called *Cameron groups* and a precise classification is known.<sup>12, 28</sup> However, this is not a new insight and the fact that Cameron groups form the bottleneck to improving Luks's method was already known in the 1980s.

**Babai's Quasipolynomial-Time Algorithm**

Attacking exactly this bottleneck, 35 years later, it was Babai who improved the running time of the theoretically fastest general graph isomorphism algorithm. He showed that graph isomorphism can be solved in quasipolynomial time  $n^{\text{poly}(\log(n))}$ , that is,  $n^{(\log(n))^c}$  for some constant  $c$ . Doing his algorithmic ideas justice is difficult not only because they span 80 pages in his original manuscript but also because the algorithm contains several major, very distinct new ideas that combine smoothly to an overall algorithm. Here, we can

Figure 6. Basic permutation group concepts.

**Basic Permutation Group Concepts**

**Permutation Domain:** The objects that are permuted.

**Symmetric group:** all permutations.

**Alternating group:** even permutations, that is, products of an even number of transpositions.

**The giants:** the symmetric and the alternating group.

**Orbits:** equivalence classes of objects that can be mapped to each other.

**Transitive group:** every object can be mapped to every other object, that is, only one orbit.

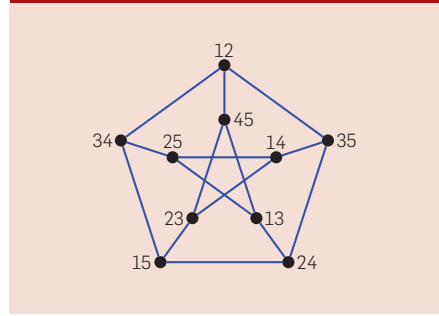
**Block:** A subset of the objects that is always mapped to itself or somewhere else entirely.

**Primitive group:** permutation group whose blocks are singletons or the entire permutation domain.

**Example:**

Vertices that are in the same orbit of the automorphism group of the graph are colored with the same color. The sets bordered by dashed lines are examples of blocks.

Figure 7. The Petersen graph is a small graph whose automorphism is a Johnson group. Its nodes correspond to the 2-element subsets of  $\{1, \dots, 5\}$ , with an edge between two nodes if the corresponding subsets have an empty intersection. The automorphism group of the Petersen graph is the symmetric group  $S_5$  with its natural action on the 2-element subsets.



thus only sketch the underlying ideas of the various puzzle pieces and how they are combined.

The first step, at quasipolynomial cost, is to reduce the bottleneck of Cameron groups further to what Babai calls *Johnson groups*. They are groups abstractly isomorphic to a symmetric or an alternating group, but do not necessarily act in their natural action of permuting elements in some ground set, and rather consist of permutations of the  $t$ -element subsets of the ground set. An example is the automorphism group of the *Petersen graph* (see Figure 7).

As next step, to make Luks's framework more flexible for his recursive algorithm, Babai does not only maintain an enclosing group  $\Gamma$  containing the automorphism group, but also a homomorphism  $\varphi: \Gamma \rightarrow \Gamma'$  into a permutation group  $\Gamma' \leq \text{Sym}(\Omega')$  over an *ideal domain*  $\Omega'$ . This allows the algorithm to make progress by decreasing the size of the ideal domain.

Initially, for Johnson groups, we can choose as ideal domain the abstract ground set mentioned here. This way, the image of  $\varphi$  contains almost all permutations, that is, it is the symmetric group or the alternating group on the ideal domain. These two groups are by far the largest primitive groups and therefore called the *giants*. Accordingly, we speak of a *giant homomorphism*.

The general strategy of the algorithm is to reduce a problem instance to quasipolynomially many instances that are all smaller than the original instance by at least a constant factor. This is continued until the recursive instances are sufficiently small to be resolved with brute force, leading to an overall quasipolynomial-time algorithm.

If the permutation group induced by the homomorphism  $\varphi$  on the ideal domain is intransitive or imprimitive, we can use the strategies of Luks to process orbits sequentially or to consider the actions on blocks, respectively. For this, some nontrivial group theory is required to pull back information from the ideal domain to the original domain.

In case the subgroup  $A$  of  $\Gamma$  that we are interested in maps onto the alternating group of the ideal domain, the computation of  $A$  is comparatively easy, so let us focus on the case that the image of the subgroup is not a giant.

We then use a *local to global* approach. We first collect local certificates by testing, for all logarithmic-size subsets  $T$  of the ideal domain, whether the homomorphism  $\varphi$  applied to the sought-after automorphism group  $A$  and restricted to  $T$  is a giant homomorphism. We call these sets *test sets*. A test set is *full* if the said restriction is a giant homomorphism. As the test set size is only logarithmic and there are only quasipolynomially many such test sets, we can test all test sets for fullness using recursion.

If a test set is full, which certifies high local symmetry, there must be global symmetries certifying this and quite surprisingly such global symmetries can be efficiently constructed. At the core of this statement lies the *Unaffected Stabiliser Lemma*, a central insight proven by Babai.

If there are a lot of full test sets, the global symmetries allow for efficient recursion. On the other hand, if only few test sets are full, the graph must have a nontrivial structural invariant. Furthermore, we can use the logarithmic-dimensional Weisfeiler-Leman algorithm to construct such a structural invariant in the form of a relational structure of logarithmic arity. This breaks the apparent symmetry.

With the *design lemma*, we can reduce the relational structure of logarithmic arity to a structure with a binary relation. We obtain a uniprimitive coherent configuration, a particular structure important in algebraic graph theory closely related to the 2-dimensional Weisfeiler-Leman algorithm.

The final puzzle piece is the *Split-Or-Johnson* combinatorial partitioning algorithm which, from a uniprimitive coherent configuration either produces a split or finds a large canonically embedded Johnson graph, a graph whose automorphism group is a Johnson group. In fact splits, which are invariant partitions of the ideal domain akin to the blocks of a permutation group, can also occur during other parts of the algorithm. They are handled with the techniques similar to the imprimitive case of Luks's algorithm.

We are left with the case in which a large canonically embedded Johnson graph has been produced. After all, this case had to occur at some point because we know that the resilient Johnson

groups exist. But now the Johnson graph is in fact a blessing because we can exploit the well-understood structure of the Johnson graphs to dramatically decrease the size of the ideal domain.

Overall we obtain a quasipolynomial-time algorithm solving the general graph isomorphism problem. Besides the original manuscript, there is also a detailed explanation of the algorithm in the Bourbaki series by Helfgott (see Helfgott et al.<sup>20</sup> for an English translation). In fact, Helfgott detected an error in the Split-or-Johnson routine which however was quickly fixed by Babai.

Babai's algorithm depends on the classification of the finite simple groups, an enormous theorem spanning several hundred journal articles written by numerous authors. Many group theorists prefer to avoid the theorem and indeed Pyber modified Babai's algorithm to give an alternative analysis that does not depend on the classification.

In further advances, Babai recently extended his result to a canonization algorithm that runs in quasipolynomial time,<sup>4</sup> and there is an improvement on Luks's original result for graphs of maximum degree  $d$  testing isomorphism in time  $n^{(\log(d)^d)}$ .<sup>18</sup>

## Practical Graph Isomorphism

In practice it is excessive to even run the 2-dimensional Weisfeiler-Leman algorithm, let alone some version of increasing dimension as in Babai's algorithm. Current isomorphism packages rather use color refinement, that is, the 1-dimensional version. As mentioned, this is already sufficient for almost all graphs. If it turns out not to be sufficient, the algorithms take the route of branching by using the concept of individualization.

Specifically, the *individualization-refinement* paradigm, which is adopted by virtually all modern competitive isomorphism tools, one by one artificially assigns a different color to all the vertices in a color class. This breaks the symmetry and subsequently color refinement can be potentially applied again to produce a more refined partition of the vertices. In a backtracking manner, the tools continue until a discrete color (a coloring in which all color classes are a singleton) has been reached. The tools use various pruning techniques, such as invariants and pruning with



automorphisms, discovered with intricate methods, to drastically improve their performance.

The tools actually compute a canonical form, which also solves the isomorphism problem (as explained earlier). This highly practical method was originally pioneered by McKay with his famous software tool Nauty. There are now various extremely efficient packages such as Bliss, Conauto, Nauty, Saucy, and Traces freely available. Recently many new ideas, responsible for their efficiency, such as the use of the trace for early abortion of color refinement in Traces, have found their way into the tools. We refer the reader to an extensive survey.<sup>31</sup> In contrast to Babai's quasipolynomial-time algorithm, there are, however, graphs on which the running time of all individualization-refinement algorithms scale exponentially.<sup>33</sup>

## Applications

Graph isomorphism tools can in practice be used to find symmetries of combinatorial objects and as such they have numerous applications in miscellaneous domains. In the context of optimisation, for example, in SAT solving, symmetries are exploited to collapse the search space, as parts equivalent under symmetries only need to be explored once. An alternative way of exploiting symmetries is to add symmetry breaking constraints to the original input again drastically improving performance.

Another application domain exploits canonical labeling to store graph structured data in a database. For example, when molecules are stored in a chemical database, the idea is to store only a canonical representative. To look up a given molecule in the database, we compute its canonical representative and find the result in the database. This way, no isomorphism tests against the elements in the database are required. Other application domains include machine learning, computer graphics, software verification, model checking, and mathematical programming.

## Concluding Remarks

With Babai's quasipolynomial-time algorithm, we have seen a breakthrough on one of the oldest and best studied algorithmic problems. Undoubtedly, this algorithm and its underlying mathematical framework rank among

the most important contributions to theoretical computer science in a long time. We are only starting to explore the potential of the wealth of new ideas they bring to the field.

Current challenges include the group isomorphism problem, one of the core obstacles to even faster graph isomorphism tests. On the practical side, emerging applications in areas such as machine learning demand a better understanding of approximate versions of isomorphism and similarity measures between graphs.

Yet the question whether graph isomorphism is solvable in polynomial time remains open, and we can expect further deep, exciting insights until it will finally be settled. **C**

## References

- Atserias, A., Maneva, E. Sherati-Adams relaxations and indistinguishability in counting logics. *SIAM J. Comput.* 1, 42 (2013), 112–137.
- Babai, L. Technical Report D.M.S. No. 79-10. *Monte Carlo Algorithms in Graph Isomorphism Testing*. Université de Montréal, 1979.
- Babai, L. Graph isomorphism in quasipolynomial time. In *Proceedings of the 48<sup>th</sup> Annual ACM Symposium on Theory of Computing (STOC'16)*, 2016), 684–697.
- Babai, L. Canonical form for graphs in quasipolynomial time. In *Proceedings of the 51<sup>st</sup> Annual ACM SIGACT Symposium on Theory of Computing* (2019).
- Babai, L., Cameron, P.J., Pálffy, P.P. On the orders of primitive groups with restricted nonabelian composition factors. *J. Algebra* 1, 79 (1982), 161–168. [https://doi.org/10.1016/0021-8693\(82\)90323-4](https://doi.org/10.1016/0021-8693(82)90323-4)
- Babai, L., Chen, X., Sun, X., Teng, S.-H., Wilmes, J. Faster canonical forms for strongly regular graphs. In *Proceedings of the 54<sup>th</sup> Annual IEEE Symposium on Foundations of Computer Science* (2013), 157–166.
- Babai, L., Codenotti, P., Grochow, J., Qiao, Y. Code equivalence and group isomorphism. In *Proceedings of the 22<sup>nd</sup> annual ACM-SIAM Symposium on Discrete Algorithms* (2011), 1395–1408.
- Babai, L., Erdős, P., Selkow, S. Random graph isomorphism. *SIAM J. Comput.* 9 (1980), 628–635.
- Babai, L., Kantor, W.M., Luks, E.M. Computational complexity and the classification of finite simple groups. In *Proceedings of the 24<sup>th</sup> Annual Symposium on Foundations of Computer Science* (1983), 162–171.
- Babai, L., Luks, E.M. Canonical labeling of graphs. In *Proceedings of the 15<sup>th</sup> ACM Symposium on Theory of Computing* (1983), 171–183.
- Cai, J., Fürer, M., Immerman, N. An optimal lower bound on the number of variables for graph identification. *Combinatorica* 12 (1992), 389–410.
- Cameron, P.J. Finite permutation groups and finite simple groups. *Bull. Lond. Math. Soc.* 13, 1 (1981), 1–22.
- Cardon, A., Crochemore, M. Partitioning a graph in  $O(|A| \log_2 |V|)$ . *Theor. Comput. Sci.* 19, 1 (1982), 85–98.
- Datta, S., Limaye, N., Nimbhorkar, P., Thierauf, T., Wagner, F. Planar graph isomorphism is in log-space. In *Proceedings of the 24<sup>th</sup> Annual IEEE Conference on Computational Complexity* (2009), 203–214.
- Garey, M.R., Johnson, D.S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, 1979.
- Goldreich, O., Micali, S., Wigderson, A. Proofs that yield nothing but their validity and a methodology of cryptographic protocol design. In *Proceedings of the 27<sup>th</sup> Annual Symposium on Foundations of Computer Science* (1986), 174–187.
- Grohe, M. Descriptive complexity, canonisation, and definable graph structure theory. *Lecture Notes in Logic*, Vol. 47. Cambridge University Press, 2017.
- Grohe, M., Neuen, D., Schweitzer, P. A faster isomorphism test for graphs of small degree. In *Proceedings of the 59<sup>th</sup> Annual IEEE Symposium on Foundations of Computer Science* (2018), 89–100.
- Grohe, M., Marx, D. Structure theorem and isomorphism test for graphs with excluded topological subgraphs. *SIAM J. Comput.* 1, 44 (2015), 114–159.
- Helfgott, H.A., Bajpai, J., Dona, D. Graph isomorphisms in quasi-polynomial time. *ArXiv* 1710.04574 (2017).
- Hopcroft, J.E. An  $n \log n$  algorithm for minimizing states in a finite automaton. In *Theory of Machines and Computations*, Z. Kohavi and A. Paz, eds. Academic Press, 1971, 189–196.
- Hopcroft, J.E., Tarjan, R. Isomorphism of planar graphs (working paper). In *Complexity of Computer Computations*, R. E. Miller and J. W. Thatcher, eds. Plenum Press, 1972.
- Karp, R.M. Reducibilities among combinatorial problems. In *Complexity of Computer Computations*, R.E. Miller and J.W. Thatcher, eds. Plenum Press, New York, 1972, 85–103.
- Kiefer, S., Ponomarenko, I., Schweitzer, P. The Weisfeiler-Leman dimension of planar graphs is at most 3. In *Proceedings of the 32<sup>nd</sup> ACM-IEEE Symposium on Logic in Computer Science* (2017).
- Lokshantov, D., Pilipczuk, M., Pilipczuk, M., Saurabh, S. Fixed-parameter tractable canonization and isomorphism test for graphs of bounded treewidth. In *Proceedings of the 55<sup>th</sup> Annual IEEE Symposium on Foundations of Computer Science* (2014), 186–195.
- Luks, E.M. Isomorphism of graphs of bounded valence can be tested in polynomial time. *J. Comput. Syst. Sci.* 25 (1982), 42–65.
- Luks, E.M. Permutation groups and polynomial-time computation. In *Groups And Computation, Proceedings of a DIMACS Workshop, New Brunswick, New Jersey, USA, October 7–10, 1991 (DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 11)*, L. Finkelstein and W.M. Kantor, eds. DIMACS/AMS, 1991, 139–176.
- Maróti, A. On the orders of primitive groups. *J. Algebra* 258, 2 (2002), 631–640.
- Mathon, R. A note on the graph isomorphism counting problem. *Inform. Process. Lett.* 8, 3 (1979), 131–132.
- McKay, B. 1981. Practical graph isomorphism. *Congr. Numer.* 30 (1981), 45–87.
- McKay B.D., Piperno, A. Practical graph isomorphism, II. *J. Symbol. Comput.* 60 (2014), 94–112.
- Morgan, H.L. The generation of a unique machine description for chemical structures—A technique developed at chemical abstracts service. *J. Chem. Document.* 5, 2 (1965), 107–113.
- Neuen, D., Schweitzer, P. An exponential lower bound for individualization-refinement algorithms for graph isomorphism. In *Proceedings of the 50<sup>th</sup> Annual ACM SIGACT Symposium on Theory of Computing* (2018), 138–150.
- Ponomarenko, I.N. The isomorphism problem for classes of graphs that are invariant with respect to contraction. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* 174, Teor. Slozhn. Vychisl. 3 (1988), 147–177, 182. (in Russian).
- Ray, L.C., Kirsch, R.A. Finding chemical records by digital computers. *Science* 126 (1957).
- Schweitzer, P., Wiebking, D. A unifying method for the design of algorithms canonizing combinatorial objects. In *Proceedings of the 51<sup>st</sup> Annual ACM SIGACT Symposium on Theory of Computing* (2019), 1247–1258.
- Ákos Seress. *Permutation group algorithms*. Cambridge Tracts in Mathematics, Vol. 152. Cambridge University Press, Cambridge, 2003, x+264 pages. <https://doi.org/10.1017/CBO9780511546549>
- Torán, J. On the hardness of graph isomorphism. *SIAM J. Comput.* 33, 5 (2004), 1093–1108.
- Unger, S. GIT—A heuristic program for testing pairs of directed line graphs for isomorphism. *Commun. ACM* 7, 1 (1964), 26–34.
- Weisfeiler, B., Leman, A. The reduction of a graph to canonical form and the algebra which appears therein. *NTI, Series 2* (1968). English translation by G. Ryabov. Available at: [https://www.itu.cz/wl2018/pdf/wl\\_paper\\_translation.pdf](https://www.itu.cz/wl2018/pdf/wl_paper_translation.pdf).

**Martin Grohe** is a professor of computer science at RWTH Aachen University, Aachen, Germany.

**Pascal Schweitzer** is a professor at TU Kaiserslautern, Germany.

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

# Introducing *ACM Transactions on Human-Robot Interaction*

Now accepting submissions to ACM THRI

In January 2018, the *Journal of Human-Robot Interaction* (JHRI) became an ACM publication and was rebranded as the *ACM Transactions on Human-Robot Interaction* (THRI). It will continue to be open access, fostering the widest possible readership of HRI research and information. All issues will be available in the ACM Digital Library.

ACM THRI aims to be the leading peer-reviewed interdisciplinary journal of human-robot interaction. Publication preference is given to articles that contribute to the state of the art or advance general knowledge, have broad interest, and are written to be intelligible to a wide range of audiences. Submitted articles must achieve a high standard of scholarship. Accepted papers must: (1) advance understanding in the field of human-robot interaction, (2) add state-of-the-art or general information to this field, or (3) challenge existing understandings in this area of research.

ACM THRI encourages submission of well-written papers from all fields, including robotics, computer science, engineering, design, and the behavioral and social sciences. Published scholarly papers can address topics including how people interact with robots and robotic technologies, how to improve these interactions and make new kinds of interaction possible, and the effects of such interactions on organizations or society. The editors are also interested in receiving proposals for special issues on particular technical problems or that leverage research in HRI to advance other areas such as social computing, consumer behavior, health, and education.

The inaugural issue of the rebranded *ACM Transactions on Human-Robot Interaction* has been published and can be found in the ACM Digital Library.

For further information and to submit your manuscript visit [thri.acm.org](http://thri.acm.org).



Association for  
Computing Machinery





## ACM BOOKS Collection II

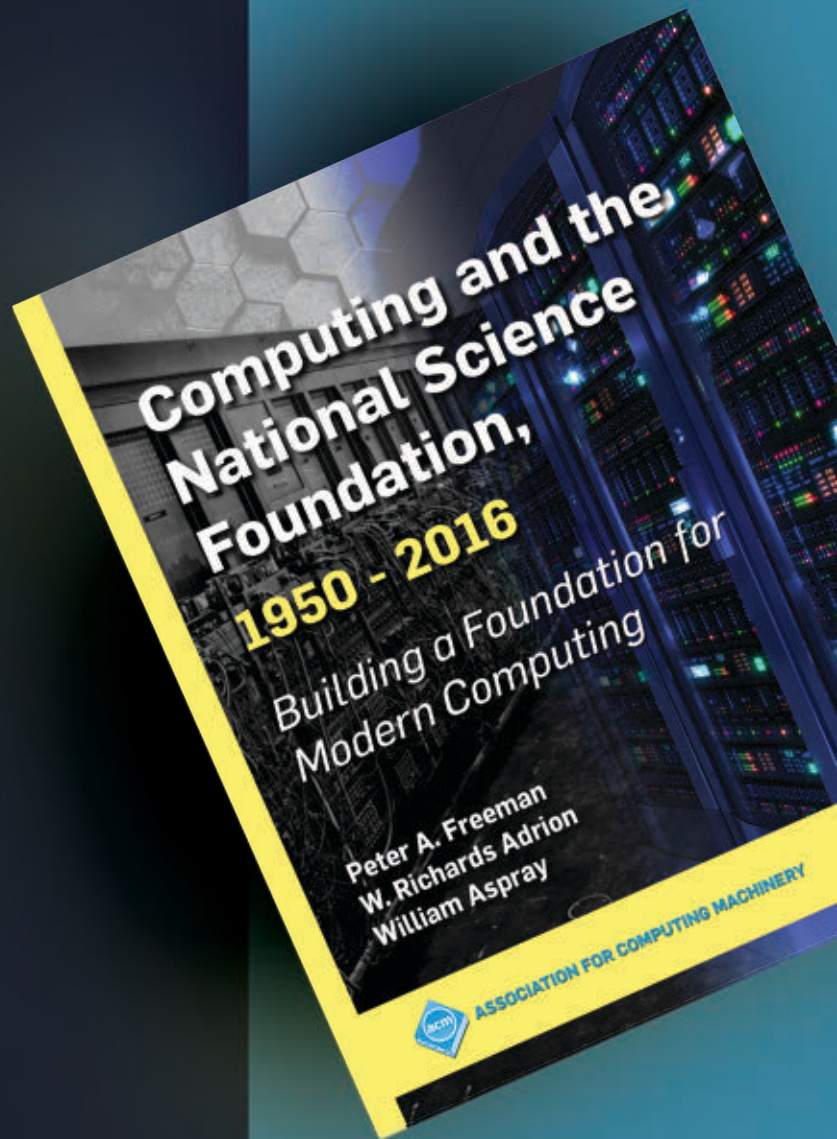
This organizational history relates the role of the National Science Foundation (NSF) in the development of modern computing. Drawing upon new and existing oral histories, extensive use of NSF documents, and the experience of two of the authors as senior managers, this book describes how NSF's programmatic activities originated and evolved to become the primary source of funding for fundamental research in computing and information technologies.

The book traces how NSF's support has provided facilities and education for computing usage by all scientific disciplines, aided in institution and professional community building, supported fundamental research in computer science and allied disciplines, and led the efforts to broaden participation in computing by all segments of society.

Today, the research and infrastructure facilitated by NSF computing programs are significant economic drivers of American society and industry. The NSF has advanced the development of human capital and ideas for future advances in computing and its applications.

This account is the first comprehensive coverage of NSF's role in the extraordinary growth and expansion of modern computing and its use. It will appeal to historians of computing, policy makers and leaders in government and academia, and individuals interested in the history and development of computing and the NSF.

<http://books.acm.org>  
<http://store.morganclaypool.com/acm>



### **Computing and the National Science Foundation 1950-2016** *Building a Foundation for Modern Computing*

**Peter A. Freeman  
W. Richards Adrion  
William Aspray**

ISBN: 978-1-4503-7271-8  
DOI: 10.1145/3335772

# research highlights

---

P. 138

**Technical  
Perspective**  
**When the Adversary  
Is Your Friend**

By Alexei A. Efros  
and Aaron Hertzmann

P. 139

## Generative Adversarial Networks

By Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu,  
David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio

---

P. 145

**Technical  
Perspective**  
**BLeak:  
Semantics-Aware  
Leak Detection  
in the Web**

By Harry Xu

P. 146

## BLeak: Automatically Debugging Memory Leaks in Web Applications

By John Vilks and Emery D. Berger



# Technical Perspective

## When the Adversary Is Your Friend

By Alexei A. Efros and Aaron Hertzmann

MOST FUNDAMENTAL IDEAS in convolutional neural networks (rebranded in 2010s as deep learning), are actually several decades old. It just took a while for the hardware, the data, and the research community to catch up. But if one asks, what is the most important *new* idea to have come out in the last decade, without a doubt, it is Generative Adversarial Networks (GANs). Like most good papers, it certainly had some precursors, yet, when it came out in 2014, there was a palpable sense that something new and exciting is afoot. After all, the paper was easy to like as it had all the right ingredients: a clever idea, nice math, an intriguing connection to evolution. And if the original paper didn't dazzle with the visual quality of its results, the long string of followup works have shown the impressive power of the method, one that may have considerable impact beyond computing.

Most of the recent successes in machine learning has come from so-called *discriminative models*: given some input data, such as an image, these models try to look for the relevant bits and pieces of information to decide what it is. For example, the presence of stripes might suggest that an image contains a zebra. An alternative are *generative models*, which aim to approximate the process that generates the data. While a discriminative model would only tell you that something is a zebra, a generative model could actually paint you one.

However, generative models have not been very successful for real-world imagery, largely because it is difficult to automatically evaluate the generator. If we had a way to measure how good a model's output is—known as an objective function or a “loss function”—we could optimize our generative model according to this metric. But how do you quantify whether a model does a good job at



Mario Klingemann, *Do Not Kill the Messenger* (2017); <https://bit.ly/3iYhvxU>

generating realistic new images that no one has ever seen before? The key insight of the following GAN paper is to learn the loss function at the same time as learning the generative model. This idea of simultaneously learning a *generator* and a *discriminator* in an adversarial manner has turned out to be extremely powerful. The model leads to vivid anthropomorphic analogies: some researchers explain GANs as a competition between two actors, like an artist and a critic, a student and a teacher, or a forger and a detective.

Upon initial publication, this paper led to dizzyingly fast advances in the quality and generality of GAN models; within a few years, researchers demonstrated the ability to generate seemingly infinite sets of new images that were virtually indistinguishable from the real thing. Moreover, learned adversarial losses turned out to be very useful in many other contexts, for example, provid-

ing “training wheels” for image editing that keep images realistic during the editing process.

GAN-based models could soon have considerable cultural and political impact on society, both positive and negative. Many notable artists, including Sofia Crespo, Scott Eaton, Mario Klingemann, Trevor Paglen, Jason Salavon, and Helena Sarin, have used GANs, and GAN art has appeared in several major galleries, festivals, and auction houses.<sup>1,2</sup> In fact, some of the power of GANs as artistic tools can be experienced using Joel Simon's Artbreeder.com website. Many movie studios and startups are currently exploring technologies using GAN losses to create virtual characters, avatars, and sets, to provide new artistic tools for storytelling and communication. GANs could help us take better pictures and capture memories of the world in 3D, and perhaps someday our video teleconferencing will be improved by GANs that render us as realistic or as fanciful avatars in shared virtual spaces. At the same time, GAN-based techniques pose major concerns around misinformation and various malicious uses of DeepFakes, as well as various data biases in image synthesis algorithms and how they are used. In addition to being an important fundamental contribution to computing, GANs are at the vanguard of some of our hopes and fears for how imaging algorithms can transform society. 

### References

1. Bailey, J. The tools of generative art, from Flash to neural networks. *Art in America* 8 (Jan. 2020); <https://bit.ly/2EQqna9>
2. Hertzmann, A. Visual Indeterminacy in GAN Art. *Leonardo* 53, 4 (Aug. 2020), 424–428; <https://bit.ly/3U2KkKa>

**Alexei A. Efros** is a professor in the EECS Department at UC Berkeley, where he is part of the Berkeley Artificial Intelligence Research Lab, Berkeley, CA, USA.

**Aaron Hertzmann** is a principal scientist at Adobe in San Francisco, CA, USA.

Copyright held by authors/owners.

# Generative Adversarial Networks

By Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu,  
David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio

## Abstract

**Generative adversarial networks are a kind of artificial intelligence algorithm designed to solve the *generative modeling* problem. The goal of a generative model is to study a collection of training examples and learn the probability distribution that generated them. Generative Adversarial Networks (GANs) are then able to generate more examples from the estimated probability distribution. Generative models based on deep learning are common, but GANs are among the most successful generative models (especially in terms of their ability to generate realistic high-resolution images). GANs have been successfully applied to a wide variety of tasks (mostly in research settings) but continue to present unique challenges and research opportunities because they are based on game theory while most other approaches to generative modeling are based on optimization.**

## 1. INTRODUCTION

Most current approaches to developing artificial intelligence are based primarily on machine learning. The most widely used and successful form of machine learning to date is supervised learning. Supervised learning algorithms are given a dataset of pairs of example inputs and example outputs. They learn to associate each input with each output and thus learning a mapping from input to output examples. The input examples are typically complicated data objects like images, natural language sentences, or audio waveforms, while the output examples are often relatively simple. The most common kind of supervised learning is classification, where the output is just an integer code identifying a specific category (a photo might be recognized as coming from category 0 containing cats, or category 1 containing dogs, etc.).

Supervised learning is often able to achieve greater than human accuracy after the training process is complete, and thus has been integrated into many products and services. Unfortunately, the learning process itself still falls far short of human abilities. Supervised learning by definition relies on a human supervisor to provide an output example for each input example. Worse, existing approaches to supervised learning often require millions of training examples to exceed human performance, when a human might be able to learn to perform the task acceptably from a very small number of examples.

In order to reduce both the amount of human supervision required for learning and the number of examples required for learning, many researchers today study *unsupervised learning*, often using *generative models*. In this overview paper, we describe one particular approach to unsupervised learning via generative modeling called *generative adversarial networks*. We briefly review

applications of GANs and identify core research problems related to convergence in games necessary to make GANs a reliable technology.

## 2. GENERATIVE MODELING

The goal of supervised learning is relatively straightforward to specify, and all supervised learning algorithms have essentially the same goal: learn to accurately associate new input examples with the correct outputs. For instance, an object recognition algorithm may associate a photo of a dog with some kind of DOG category identifier.

Unsupervised learning is a less clearly defined branch of machine learning, with many different unsupervised learning algorithms pursuing many different goals. Broadly speaking, the goal of unsupervised learning is to learn something useful by examining a dataset containing unlabeled input examples. Clustering and dimensionality reduction are common examples of unsupervised learning.

Another approach to unsupervised learning is generative modeling. In generative modeling, training examples  $\mathbf{x}$  are drawn from an unknown distribution  $p_{\text{data}}(\mathbf{x})$ . The goal of a generative modeling algorithm is to learn a  $p_{\text{model}}(\mathbf{x})$  that approximates  $p_{\text{data}}(\mathbf{x})$  as closely as possible.

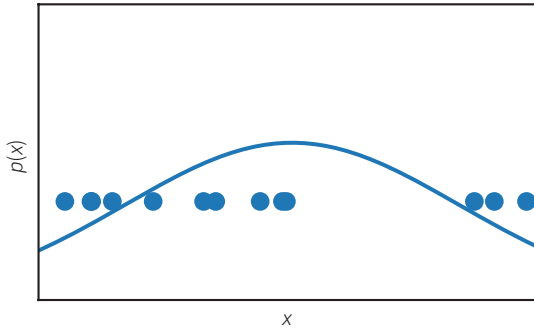
A straightforward way to learn an approximation of  $p_{\text{data}}$  is to explicitly write a function  $p_{\text{model}}(\mathbf{x}; \theta)$  controlled by parameters  $\theta$  and search for the value of the parameters that makes  $p_{\text{data}}$  and  $p_{\text{model}}$  as similar as possible. In particular, the most popular approach to generative modeling is probably *maximum likelihood estimation*, consisting of minimizing the Kullback-Leibler divergence between  $p_{\text{data}}$  and  $p_{\text{model}}$ . The common approach of estimating the mean parameter of a Gaussian distribution by taking the mean of a set of observations is one example of maximum likelihood estimation. This approach based on explicit density functions is illustrated in Figure 1.

Explicit density modeling has worked well for traditional statistics, using simple functional forms of probability distributions, usually applied to small numbers of variables. More recently, with the rise of machine learning in general and deep learning in particular, researchers have become interested in learning models that make use of relatively complicated functional forms. When a deep neural network is used to generate data, the corresponding density function may be computationally intractable. Traditionally, there have been two dominant approaches to confronting this intractability problem: (1) carefully design the model to have a tractable density function (e.g., Frey<sup>11</sup>) and (2) design a learning algorithm based on

The original version of this paper is entitled “Generative Adversarial Networks” and was published in *Advances in Neural Information Processing Systems 27* (NIPS 2014).



**Figure 1.** Many approaches to generative modeling are based on *density estimation*: observing several training examples of a random variable  $x$  and inferring a density function  $p(x)$  that generates the training data. This approach is illustrated here, with several data points on a real number line used to fit a Gaussian density function that explains the observed samples. In contrast to this common approach, GANs are *implicit models* that infer the probability distribution  $p(x)$  without necessarily representing the density function explicitly.



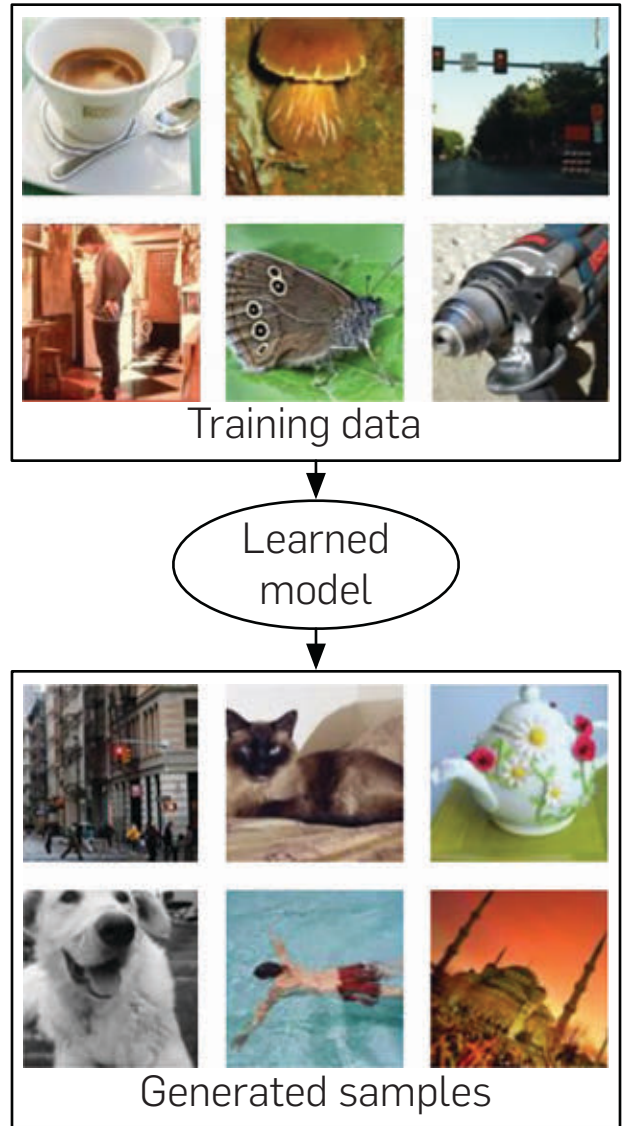
a computationally tractable approximation of an intractable density function (e.g., Kingma and Welling<sup>15</sup>). Both approaches have proved difficult, and for many applications, such as generating realistic high resolution images, researchers remain unsatisfied with the results so far. This motivates further research to improve these two paths, but also suggests that a third path could be useful.

Besides taking a point  $x$  as input and returning an estimate of the probability of generating that point, a generative model can be useful if it is able to generate a sample from the distribution  $p_{\text{model}}$ . This is illustrated in Figure 2. Many models that represent a density function can also generate samples from that density function. In some cases, generating samples is very expensive or only approximate methods of generating samples are tractable.

Some generative models avoid the entire issue of designing a tractable density function and learn only a tractable sample generation process. These are called *implicit generative models*. GANs fall into this category. Prior to the introduction of GANs, the state of the art deep implicit generative model was the *generative stochastic network*<sup>4</sup> which is capable of approximately generating samples via an incremental process based on Markov chains. GANs were introduced in order to create a deep implicit generative model that was able to generate true samples from the model distribution in a single generation step, without need for the incremental generation process or approximate nature of sampling Markov chains.

Today, the most popular approaches to generative modeling are probably GANs, variational autoencoders,<sup>15</sup> and fully-visible belief nets (e.g., Frey<sup>11, 26</sup>). None of these approaches relies on Markov chains, so the reason for the interest in GANs today is not that they succeeded at their original goal of generative modeling without Markov chains, but rather that they have succeeded in generating high-quality images and have proven useful for several tasks other than straightforward generation, as described in Section 5.

**Figure 2.** The goal of many generative models, as illustrated here, is to study a collection of training examples, then learn to generate more examples that come from the same probability distribution. GANs learn to do this without using an explicit representation of the density function. One advantage of the GAN framework is that it may be applied to models for which the density function is computationally intractable. The samples shown here are all samples from the ImageNet dataset,<sup>8</sup> including the ones labeled “model samples.” We use actual ImageNet data to illustrate the goal that a hypothetical perfect model would attain.



### 3. GENERATIVE ADVERSARIAL NETWORKS

Generative adversarial networks are based on a game, in the sense of game theory, between two machine learning models, typically implemented using neural networks.

One network called the *generator* defines  $p_{\text{model}}(x)$  implicitly. The generator is not necessarily able to evaluate the density function  $p_{\text{model}}$ . For some variants of GANs, evaluation of the density function is possible (any tractable density model for which sampling is tractable and differentiable could be trained as a GAN generator, as done by Danihelka

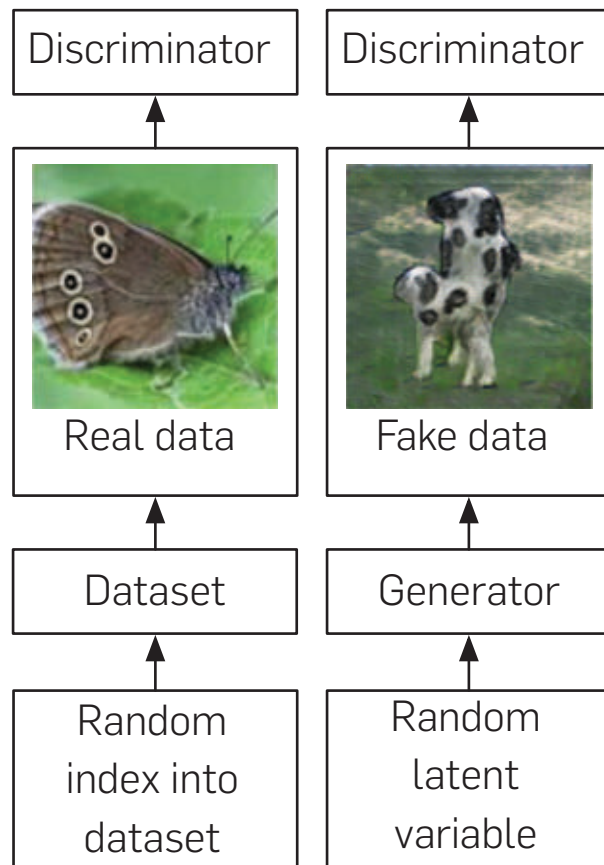
et al.<sup>6</sup>), but this is not required. Instead, the generator is able to draw samples from the distribution  $p_{\text{model}}$ . The generator is defined by a prior distribution  $p(z)$  over a vector  $z$  that serves as input to the *generator function*  $G(z; \theta^{(G)})$  where  $\theta^{(G)}$  is a set of learnable parameters defining the generator's strategy in the game. The input vector  $z$  can be thought of as a source of randomness in an otherwise deterministic system, analogous to the seed of pseudorandom number generator. The prior distribution  $p(z)$  is typically a relatively unstructured distribution, such as a high-dimensional Gaussian distribution or a uniform distribution over a hypercube. Samples  $z$  from this distribution are then just noise. The main role of the generator is to learn the function  $G(z)$  that transforms such unstructured noise  $z$  into realistic samples.

The other player in this game is the *discriminator*. The discriminator examines samples  $x$  and returns some estimate  $D(x; \theta^{(D)})$  of whether  $x$  is real (drawn from the training distribution) or fake (drawn from  $p_{\text{model}}$  by running the generator). In the original formulation of GANs, this estimate consists of a probability that the input is real rather than fake assuming that the real distribution and fake distribution are sampled equally often. Other formulations (e.g., Arjovsky et al.<sup>1</sup>) exist but generally speaking, at the level of verbal, intuitive descriptions, the discriminator tries to predict whether the input was real or fake.

Each player incurs a cost:  $J^{(G)}(\theta^{(G)}, \theta^{(D)})$  for the generator and  $J^{(D)}(\theta^{(G)}, \theta^{(D)})$  for the discriminator. Each player attempts to minimize its own cost. Roughly speaking, the discriminator's cost encourages it to correctly classify data as real or fake, while the generator's cost encourages it to generate samples that the discriminator incorrectly classifies as real. Very many different specific formulations of these costs are possible and so far most popular formulations seem to perform roughly the same.<sup>18</sup> In the original version of GANs,  $J^{(D)}$  was defined to be the negative log-likelihood that the discriminator assigns to the real-vs-fake labels given the input to the discriminator. In other words, the discriminator is trained just like a regular binary classifier. The original work on GANs offered two versions of the cost for the generator. One version, today called *minimax GAN* (M-GAN) defined a cost  $J^{(G)} = -J^{(D)}$ , yielding a minimax game that is straightforward to analyze theoretically. M-GAN defines the cost for the generator by flipping the sign of the discriminator's cost; another approach is the *non-saturating GAN* (NS-GAN), for which the generator's cost is defined by flipping the discriminator's *labels*. In other words, the generator is tried to minimize the negative log-likelihood that the discriminator assigns to the *wrong* labels. The later helps to avoid gradient saturation while training the model.

We can think of GANs as a bit like counterfeiters and police: the counterfeiters make fake money while the police try to arrest counterfeiters and continue to allow the spending of legitimate money. Competition between counterfeiters and police leads to more and more realistic counterfeit money until eventually the counterfeiters produce perfect fakes and the police cannot tell the difference between real and fake money. One complication to this analogy is that the generator learns via the discriminator's

**Figure 3. Training GANs involves training both a generator network and a discriminator network. The process involves both real data drawn from a dataset and fake data created continuously by the generator throughout the training process. The discriminator is trained much like any other classifier defined by a deep neural network. As shown on the left, the discriminator is shown data from the training set. In this case, the discriminator is trained to assign data to the “real” class. As shown on the right, the training process also involves fake data. The fake data is constructed by first sampling a random vector  $z$  from a prior distribution over latent variables of the model. The generator is then used to produce a sample  $x = G(z)$ . The function  $G$  is simply a function represented by a neural network that transforms the random, unstructured  $z$  vector into structured data, intended to be statistically indistinguishable from the training data. The discriminator then classifies this fake data. The discriminator is trained to assign this data to the “fake” class. The backpropagation algorithm makes it possible to use the derivatives of the discriminator's output with respect to the discriminator's input to train the generator. The generator is trained to fool the discriminator, in other words, to make the discriminator assign its input to the “real” class. The training process for the discriminator is thus much the same as for any other binary classifier with the exception that the data for the “fake” class comes from a distribution that changes constantly as the generator learns rather than from a fixed distribution. The learning process for the generator is somewhat unique, because it is not given specific targets for its output, but rather simply given a reward for producing outputs that fool its (constantly changing) opponent.**

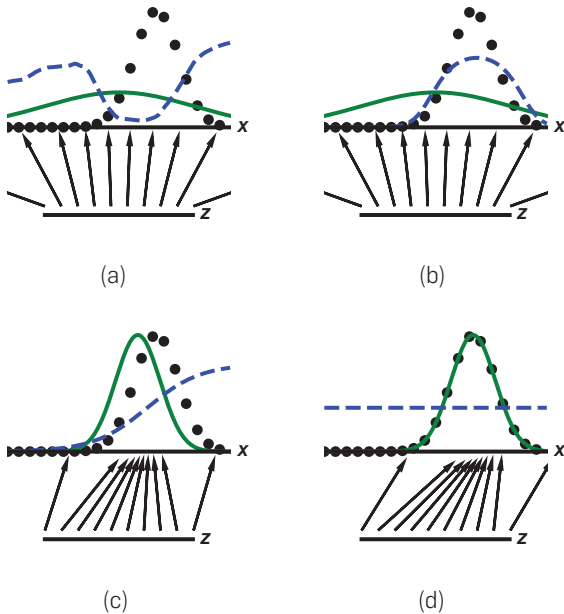


gradient, as if the counterfeiters have a mole among the police reporting the specific methods that the police use to detect fakes.

This process is illustrated in Figure 3. Figure 4 shows a cartoon giving some intuition for how the process works.



**Figure 4.** An illustration of the basic intuition behind the GAN training process, illustrated by fitting a 1-D Gaussian distribution. In this example, we can understand the goal of the generator as learning a simple scaling of the inverse cumulative distribution function of the data generating distribution. GANs are trained by simultaneously updating the discriminator function ( $D$ , blue, dashed line) so that it discriminates between samples from the data generating distribution (black, dotted line)  $p_x$  from those of the generative distribution  $p_{\text{model}}$  (green, solid line). The lower horizontal line is the domain from which  $z$  is sampled, in this case uniformly. The horizontal line above is part of the domain of  $x$ . The upward arrows show how the mapping  $x = G(z)$  imposes the non-uniform distribution  $p_{\text{model}}$  on transformed samples.  $G$  contracts in regions of high density and expands in regions of low density of  $p_{\text{model}}$ . (a) Consider a pair of adversarial networks at initialization:  $p_{\text{model}}$  is initialized to a unit Gaussian for this example while  $D$  is defined by a randomly initialized deep neural network. (b) Suppose that  $D$  were trained to convergence while  $G$  were held fixed. In practice, both are trained simultaneously, but for the purpose of building intuition, we see that if  $G$  were fixed,  $D$  would converge to  $D^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)}$ . (c) Now suppose that we gradually train both  $G$  and  $D$  for a while. The samples  $x$  generated by  $G$  flow in the direction of increasing  $D$  in order to arrive at regions that are more likely to be classified as data. Meanwhile the estimate of  $D$  is updated in response to this update in  $G$ . (d) At the Nash equilibrium, neither player can improve its payoff because  $p_{\text{model}} = p_{\text{data}}$ . The discriminator is unable to differentiate between the two distributions, that is,  $D(x) = \frac{1}{2}$ . This constant function shows that all points are equally likely to have come from either distribution. In practice,  $G$  and  $D$  are typically optimized with simultaneous gradient steps, and it is not necessary for  $D$  to be optimal at every step as shown in this intuitive cartoon. See Refs. Fedus et al.<sup>10</sup> and Nagarajan and Kolter<sup>24</sup> for more realistic discussions of the GAN equilibration process.



The situation is not straightforward to model as an optimization problem because each player's cost is a function of the other player's parameters, but each player may control only its own parameters. It is possible to reduce the situation to optimization, where the goal is to minimize

$$J^{(G)}\left(\theta^{(G)}, \operatorname{argmin}_{\theta^{(D)}} J^{(D)}\left(\theta^{(G)}, \theta^{(D)}\right)\right),$$

as demonstrated by Metz et al.,<sup>22</sup> but the argmin operation is difficult to work with in this way. The most popular approach is to regard this situation as a game between two players. Much of the game theory literature is concerned with games that have discrete and finite action spaces, convex losses, or other properties simplifying them. GANs require use of game theory in settings that are not yet well-explored, where the costs are non-convex and the actions and policies are continuous and high-dimensional (regardless of whether we consider an action to be choosing a specific parameter vector  $\theta^{(G)}$  or whether we consider the action to be generating a sample  $x$ ). The goal of a machine learning algorithm in this context is to find a *local Nash equilibrium*<sup>28</sup>: a point that is a local minimum of each player's cost with respect to that player's parameters. With local moves, no player can reduce its cost further, assuming the other player's parameters do not change.

The most common training algorithm is simply to use a gradient-based optimizer to repeatedly take simultaneous steps on both players, incrementally minimizing each player's cost with respect to that player's parameters.

At the end of the training process, GANs are often able to produce realistic samples, even for very complicated datasets containing high-resolution images. An example is shown in Figure 5.

At a high level, one reason that the GAN framework is successful may be that it involves very little approximation. Many other approaches to generative modeling must approximate an intractable density functions. GANs do not involve any

**Figure 5.** This image is a sample from a Progressive GAN<sup>14</sup> depicting a person who does not exist but was "imagined" by a GAN after training on photos of celebrities.



**Figure 6. An illustration of progress in GAN capabilities over the course of approximately three years following the introduction of GANs. GANs have rapidly become more capable, due to changes in GAN algorithms, improvements to the underlying deep learning algorithms, and improvements to underlying deep learning software and hardware infrastructure. This rapid progress means that it is infeasible for any single document to summarize the state-of-the-art GAN capabilities or any specific set of best practices; both continue to evolve rapidly enough that any comprehensive survey quickly becomes out of date. Figure reproduced with permission from Brundage et al.<sup>5</sup> The individual results are from Refs. Goodfellow,<sup>13</sup> Karras et al.,<sup>14</sup> Liu and Tuzel,<sup>17</sup> and Radford et al.<sup>27</sup> respectively.**



approximation to their true underlying task. The only real error is the statistical error (sampling of a finite amount of training data rather than measuring the true underlying data-generating distribution) and failure of the learning algorithm to converge to exactly the optimal parameters. Many generative modeling strategies would introduce these sources of error and also further sources of approximation error, based on Markov chains, optimization of bounds on the true cost rather than the cost itself, etc.

It is difficult to give much further specific guidance regarding the details of GANs because GANs are such an active research area and most specific advice quickly becomes out of date. Figure 6 shows how quickly the capabilities of GANs have progressed in the years since their introduction.

#### 4. CONVERGENCE OF GANS

The central theoretical results presented in the original GAN paper<sup>13</sup> were that:

1. in the space of density functions  $p_{\text{model}}$  and discriminator functions  $D$ , there is only one local Nash equilibrium, where  $p_{\text{model}} = p_{\text{data}}$ .
2. if it were possible to optimize directly over such density functions, then the algorithm that consists of optimizing  $D$  to convergence in the inner loop, then making a small gradient step on  $p_{\text{model}}$  in the outer loop, converges to this Nash equilibrium.

However, the theoretical model of local moves directly in density function space may not be very relevant to GANs as they are trained in practice: using local moves in *parameter* space of the *generator* function, among the set of functions representable by neural networks with a finite number of parameters, with each parameter represented with a finite number of bits.

In many different theoretical models, it is interesting to study whether a Nash equilibrium exists,<sup>2</sup> whether any

spurious Nash equilibria exist,<sup>32</sup> whether the learning algorithm converges to a Nash equilibrium,<sup>24</sup> and if it does so, how quickly.<sup>21</sup>

In many cases of practical interest, these theoretical questions are open, and the best learning algorithms seem empirically to often fail to converge. Theoretical work to answer these questions is ongoing, as is work to design better costs, models, and training algorithms with better convergence properties.

#### 5. OTHER GAN TOPICS


This article is focused on a summary of the core design considerations and algorithmic properties of GANs.

Many other topics of potential interest cannot be considered here due to space consideration. This article discussed using GANs to approximate a distribution  $p(x)$  they have also been extended to the conditional setting<sup>23,25</sup> where they generate samples corresponding to some input by drawing samples from the conditional distribution  $p(x | y)$ . GANs are related to moment matching<sup>16</sup> and optimal transport.<sup>1</sup> A quirk of GANs that is made especially clear through their connection to MMD and optimal transport is that they may be used to train generative models for which  $p_{\text{model}}$  has support only on a thin manifold and may actually assign zero likelihood to the training data. GANs struggle to generate discrete data because the back-propagation algorithm needs to propagate gradients from the discriminator through the output of the generator, but this problem is being gradually resolved.<sup>9</sup> Like most generative models, GANs can be used to fill in gaps in missing data.<sup>34</sup> GANs have proven very effective for learning to classify data using very few labeled training examples.<sup>29</sup> Evaluating the performance of generative models including GANs is a difficult research area in its own right.<sup>29, 31, 32, 33</sup> GANs can be seen as a way for machine learning to learn its own cost function, rather than minimizing a hand-designed cost function. GANs can be seen as a way of supervising machine learning by asking it to



produce any output that the machine learning algorithm itself recognizes as acceptable, rather than by asking it to produce a specific example output. GANs are thus great for learning in situations where there are many possible correct answers, such as predicting the many possible futures that can happen in video generation.<sup>19</sup> GANs and GAN-like models can be used to learn to transform data from one domain into data from another domain, even without any labeled pairs of examples from those domains (e.g., Zhu et al.<sup>35</sup>). For example, after studying a collection of photos of zebras and a collection of photos of horses, GANs can turn a photo of a horse into a photo of a zebra.<sup>35</sup> GANs have been used in science to simulate experiments that would be costly to run even in traditional software simulators.<sup>7</sup> GANs can be used to create fake data to train other machine learning models, either when real data would be hard to acquire<sup>30</sup> or when there would be privacy concerns associated with real data.<sup>3</sup> GAN-like models called domain-adversarial networks can be used for domain adaptation.<sup>12</sup> GANs can be used for a variety of interactive digital media effects where the end goal is to produce compelling imagery.<sup>35</sup> GANs can even be used to solve variational inference problems used in other approaches to generative modeling.<sup>20</sup> GANs can learn useful embedding vectors and discover concepts like gender of human faces without supervision.<sup>27</sup>

## 6. CONCLUSION

GANs are a kind of generative model based on game theory. They have had great practical success in terms of generating realistic data, especially images. It is currently still difficult to train them. For GANs to become a more reliable technology, it will be necessary to design models, costs, or training algorithms for which it is possible to find good Nash equilibria consistently and quickly. 

### References

1. Arjovsky, M., Chintala, S., Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).
2. Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573* (2017).
3. Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Greene, C.S. Privacy-preserving generative deep neural networks support clinical data sharing. *bioRxiv* (2017), 159756.
4. Bengio, Y., Thibodeau-Laufer, E., Alain, G., Yosinski, J. Deep generative stochastic networks trainable by backprop. In *ICML'2014* (2014).
5. Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G.C., Steinhart, J., Flynn, C., hEigeartaigh, S.O., Beard, S., Belfield, H., Farquhar, S., Lyle, C., Crotofo, R., Evans, O., Page, M., Bryson, J., Yampolskiy, R., Amodei, D. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *ArXiv e-prints* (Feb. 2018).
6. Danihelka, I., Lakshminarayanan, B., Uria, B., Wierstra, D., Dayan, P. Comparison of maximum likelihood and GAN-based training of real nvp. *arXiv preprint arXiv:1705.05263* (2017).
7. de Oliveira, L., Paganini, M., Nachman, B. Learning particle physics by example: location-aware generative adversarial networks for physics synthesis. *Computing and Software for Big Science* 1 1(2017), 4.
8. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09* (2009).
9. Fedus, W., Goodfellow, I., Dai, A.M. MaskGAN: Better text generation via filling in the \_\_\_\_\_. In *International Conference on Learning Representations* (2018).
10. Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A.M., Mohamed, S., Goodfellow, I. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In *International Conference on Learning Representations* (2018).
11. Frey, B.J. *Graphical Models for Machine Learning and Digital Communication*. MIT Press, Boston, 1998.
12. Ganin, Y., Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning* (2015), 1180–1189.
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative adversarial nets.

- Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger, eds. *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., Boston, 2014, 2672–2680.
14. Karras, T., Aila, T., Laine, S., Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *CoRR*, abs/1710.10196 (2017).
15. Kingma, D.P., Welling, M. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2014).
16. Li, Y., Swersky, K., Zemel, R.S. Generative moment matching networks. *CoRR*, abs/1502.02761 (2015).
17. Liu, M.-Y., Tuzel, O. Coupled generative adversarial networks. D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett, eds. *Advances in Neural Information Processing Systems 29*, Curran Associates, Inc., Boston, 2016, 469–477.
18. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O. Are GANs created equal? a large-scale study. *arXiv preprint arXiv:1711.10337* (2017).
19. Mathieu, M., Couprie, C., LeCun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).
20. Mescheder, L., Nowozin, S., Geiger, A. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. *arXiv preprint arXiv:1701.04722* (2017).
21. Mescheder, L., Nowozin, S., Geiger, A. The numerics of gans. In *Advances in Neural Information Processing Systems* (2017), 1823–1833.
22. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163* (2016).
23. Mirza, M., Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).
24. Nagarajan, V., Kolter, J.Z. Gradient descent GAN optimization is locally stable. I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, eds. *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., Boston, 2017, 5585–5595.
25. Odena, A., Olah, C., Shlens, J. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585* (2016).
26. Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L.C., Stimberg, F., et al. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433* (2017).
27. Radford, A., Metz, L., Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).
28. Ratliff, L.J., Burden, and S.A., Sastry, S.S. Characterization and computation of local nash equilibria in continuous games. In *Communication, Control, and Computing (Allerton)*, 2013 51<sup>st</sup> Annual Allerton Conference on. IEEE, (2013), 917–924.
29. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems* (2016), 2234–2242.
30. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R. Learning from simulated and unsupervised images through adversarial training.
31. Theis, L., van den Oord, A., Bethge, M. A note on the evaluation of generative models. *arXiv:1511.01844* (Nov 2015).
32. Unterthiner, T., Nessler, B., Klambauer, G., Heusel, M., Ramsauer, H., Hochreiter, S. Coupled GANs: Provably optimal Nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819* (2017).
33. Wu, Y., Burda, Y., Salakhutdinov, R., Grosse, R. On the quantitative analysis of decoder-based generative models. *arXiv preprint arXiv:1611.04273* (2016).
34. Yeh, R., Chen, C., Lim, T.Y., Hasegawa-Johnson, M., Do, M.N. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539* (2016).
35. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).

Ian Goodfellow, written while at Google Brain.

Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, Université de Montréal.

Final submitted 5/9/2018.

# Technical Perspective

## BLEak: Semantics-Aware Leak Detection in the Web

By Harry Xu

WEB APPLICATIONS ARE at least as likely to leak memory as regular applications. Web leaks can significantly increase a browser's memory footprint, reducing application responsiveness and even crashing browser tabs. Such leaks exist everywhere, on websites that people use on a daily basis—Google Maps, Firefox, Google Analytics, or Airbnb, just to name a few. They are notoriously difficult to diagnose: developers see the growth of memory usage, but where exactly are the statements that cause the growth?

Despite a rich literature of leak detection for regular (Java, C++, Python, and so on) applications, prior techniques do not work well for Web applications where leaks exhibit very different characteristics. For example, the developer may forget to remove certain event listeners and hence these listener objects are still reachable in the heap. While they are no longer used by the application, they still respond to events (for example, when the user uses the mouse on the editor), keeping their states “fresh.” As a result, existing techniques that identify suspicious objects based on their staleness (that is, time since their last access)—which have worked effectively on a wide range of traditional applications—would miss these leaks in Web applications entirely.

A key research question here is: What is the right leak oracle that can precisely capture the behavior of leaks in Web applications? In other words, what kinds of objects should be considered suspicious? Once this question is answered, developing a dynamic analysis that finds such objects would be just a step away.

The following paper provides a simple and yet unexpected answer to this question: What distinguishes leaking objects from normally behaved objects is whether their be-


**With BLEak, the authors were able to precisely and quickly identify important leaks in widely used Web applications including Airbnb and Firefox debugger.**

havior obeys certain high-level semantic rules as opposed to low-level semantics-agnostic access patterns. One clear semantic rule in Web applications is that if a user navigates to a Web page and later returns to the original page, the application's memory consumption should remain (approximately) the same. In other words, the memory consumption for such navigation “round trips” can be used as a leak oracle—if the application consumes significantly more memory when coming back to the original page, the application has a high chance of leaking memory.

Based on this observation, the authors created BLEak, a Web debugger that can help developers quickly find causes of leaks. BLEak uses a user-defined script to drive an application into a loop of navigation round trips. Next, it identifies heap paths that are growing each round trip by differencing heap snapshots. BLEak ranks these paths to find “leak roots,” cap-

tures call stacks associated with top-ranked leak roots and reports them together to the user for leak diagnosis. With BLEak, the authors were able to precisely and quickly identify important leaks in widely used Web applications including Airbnb and Firefox debugger.

These results are both impressive and aspiring, particularly in the context of at least 20 years of memory leak research. Prior work uncovers a range of low-level “symptoms” that characterize leaks for a variety of applications. These symptoms are defined at the level of object read and write and often far away from actual causes of leaks. When new applications emerge, these old symptoms no longer correlate with leaks. BLEak takes a step further by exploring semantics-aware diagnosis and demonstrates that simple semantic information provided by developers (for example, round trips) can enable heap tracking that is orders of magnitude more precise than semantics-agnostic symptoms used by conventional approaches.

Looking forward, semantics-aware bug diagnosis and optimization is an exciting research direction, especially given that modern applications and workloads are becoming increasingly complex and diverse. Semantics-agnostic approaches would be either unscalable to large code bases/heaps or unable to adapt to the high diversity in modern workloads. Future work, potentially inspired by the observation made in this paper, will determine how program semantics can be employed to optimize applications in different domains. 

Harry Xu is an associate professor in the computer science department at the University of California Los Angeles, CA, USA.

Copyright held by authors/owners.



# BLEAK: Automatically Debugging Memory Leaks in Web Applications

By John Vilkk and Emery D. Berger

## Abstract

**Memory leaks in web applications are pervasive and difficult to debug. Leaks degrade responsiveness by increasing garbage collection costs and can even lead to browser tab crashes. Previous leak detection approaches designed for conventional applications are ineffective in the browser environment. Tracking down leaks currently requires intensive manual effort by web developers, which is often unsuccessful.**

**This paper introduces BLEAK (Browser Leak debugger), the first system for automatically debugging memory leaks in web applications. BLEAK’s algorithms leverage the observation that in modern web applications, users often repeatedly return to the same (approximate) visual state (e.g., the inbox view in Gmail). Sustained growth between round trips is a strong indicator of a memory leak. To use BLEAK, a developer writes a short script (17–73 LOC on our benchmarks) to drive a web application in round trips to the same visual state. BLEAK then automatically generates a list of leaks found along with their root causes, ranked by return on investment. Guided by BLEAK, we identify and fix over 50 memory leaks in popular libraries and apps including Airbnb, AngularJS, Google Analytics, Google Maps SDK, and jQuery. BLEAK’s median precision is 100%; fixing the leaks it identifies reduces heap growth by an average of 94%, saving from 0.5MB to 8MB per round trip.**

## 1. INTRODUCTION

Browsers are one of the most popular applications on both smartphones and desktop platforms. They also have an established reputation for consuming significant amounts of memory. To address this problem, browser vendors have spent considerable effort on shrinking their browsers’ memory footprints<sup>5, 11</sup> and building tools that track the memory consumption of specific browser components.<sup>4, 10</sup>

Memory leaks in web applications only exacerbate the situation by further increasing browser memory footprints. These leaks happen when the application references unneeded state, preventing the garbage collector from collecting it. Web application memory leaks can take many forms, including failing to dispose of unneeded event listeners, repeatedly injecting iframes and CSS files, and failing to call cleanup routines in third-party libraries. Leaks are a serious concern for developers since they lead to higher garbage collection frequency and overhead. They reduce application responsiveness and can even trigger browser tab crashes by exhausting available memory.

Despite the fact that memory leaks in web applications are a well-known and pervasive problem, there are no effective automated tools that can find them. The reason is that existing memory leak detection techniques are ineffective in the browser: *leaks in web applications are fundamentally different from leaks in traditional C, C++, and Java programs.* Staleness-based techniques assume leaked memory is rarely touched,<sup>2, 6, 12, 14, 16</sup> but web applications regularly interact with leaked state (e.g., via event listeners). Growth-based techniques assume that leaked objects are uniquely owned or form strongly connected components in the heap graph.<sup>9, 16</sup> In web applications, leaked objects frequently have multiple owners, and the entire heap graph is often strongly connected due to widespread references to the global scope (window).

Faced with this lack of automated tool support, developers are currently forced to manually inspect heap snapshots to locate objects that the application incorrectly retains.<sup>1, 8</sup> Unfortunately, these snapshots do not necessarily provide actionable information (see Section 2.1). They simultaneously provide too much information (every object on the heap) and not enough information to actually debug these leaks (no connection to the code responsible for leaks). Since JavaScript is dynamically typed, most objects in snapshots are labeled as objects or arrays, which provides little assistance in locating leak sources. The result is that even expert developers are unable to find leaks: for example, a Google developer closed a Google Maps SDK leak (with 117 stars and 62 comments) because it was “infeasible” to fix as they were “not really sure in how many places [it’s] leaking”.<sup>1</sup>

We address these challenges with BLEAK (Browser Leak debugger), the first system for automatically debugging memory leaks in web applications. BLEAK leverages the following fact: over a single session, users repeatedly return to the same visual state in modern web sites, such as Facebook, Airbnb, and Gmail. For example, Facebook users repeatedly return to the news feed, Airbnb users repeatedly return to the page listing all properties in a given area, and Gmail users repeatedly return to the inbox view.

We observe that *these round trips can be viewed as an*

<sup>1</sup> <https://issuetracker.google.com/issues/35821412>.

The original version of this paper appeared in the *Proceedings of the 39<sup>th</sup> ACM SIGPLAN Conference on Programming Language Design and Implementation* (Philadelphia, PA, USA, June 18–22, 2018), 15–29.

oracle to identify leaks. Each time a web application returns to the same visual state, it should consume approximately the same amount of memory. Sustained memory growth across round trips is thus a clear indicator of a memory leak. BLEAK builds directly on this observation to find memory leaks in web applications, which (as Section 6 shows) are widespread and severe.

To use BLEAK, a developer provides a short script (17–73 LOC on our benchmarks) to drive a web application in a loop that takes round trips through a specific visual state. BLEAK then proceeds automatically, identifying memory leaks, ranking them, and locating their root cause in the source code. BLEAK first uses heap differencing to locate locations in the heap with sustained growth between each round trip, which it identifies as leak roots. To directly identify the root causes of growth, BLEAK employs JavaScript rewriting to target leak roots and collect stack traces when they grow. Finally, when presenting the results to the developer, BLEAK ranks leak roots by return on investment using a novel metric called LeakShare that prioritizes leaks that free the most memory with the least effort by dividing the “credit” for retaining a shared leaked object equally among the leak roots that retain them. This ranking focuses developer effort on the most important leaks first.

Guided by BLEAK, we identify and fix over 50 memory leaks in popular JavaScript libraries and applications including Airbnb, AngularJS, jQuery, Google Analytics, and Google Maps SDK. BLEAK has a median precision of 100% (97% on average). Its precise identification of root causes of leaks makes it relatively straightforward for us to fix nearly all of the leaks we identify (all but one). Fixing these leaks reduces heap growth by 94% on average, saving from 0.5MB to 8MB per return trip to the same visual state. We have submitted patches for all of these leaks to the application developers; at the time of writing, 16 have already been accepted and 4 are in the process of code review.

This paper makes the following contributions:

- It introduces novel techniques for automatically locating, diagnosing, and ranking memory leaks in web applications (Section 3) and presents algorithms for each (Section 4).
- It presents BLEAK, an implementation of these techniques. BLEAK’s analyses drive websites using Chrome and a proxy that transparently rewrites JavaScript code to diagnose leaks, letting it operate on unmodified websites (including over HTTPS) (Section 5).
- Using BLEAK, we identify and fix numerous leaks in widely used web applications and JavaScript libraries (Section 6).

## 2. BACKGROUND

Before presenting BLEAK and its algorithms, we first describe a representative memory leak we discovered using BLEAK (see Figure 1) and discuss why prior techniques and tools fall short when debugging leaks in web applications.

This memory leak is in Firefox’s debugger, which runs as

**Figure 1. This code from Firefox’s debugger (truncated for readability) leaks 0.5MB every time a developer opens a source file (Section 2). BLEAK finds all four leaks automatically.**

```
1 class Preview extends PureComponent {
2   // Runs when Preview is added to GUI
3   componentDidMount() {
4     const { codeMirror } = this.props.editor;
5     const wrapper = codeMirror.getWrapperElement();
6     codeMirror.on("scroll", this.onScroll);
7     wrapper.addEventListener("mouseover", this._mover);
8     wrapper.addEventListener("mouseup", this._mup);
9     wrapper.addEventListener("mousedown", this._mdown);
10  }
11 }
```

a normal web application in all browsers. Lines 6–9 register four event listeners on the debugger’s text editor (`codeMirror`) and its GUI object (`wrapper`) every time the user views a source file. The leak occurs because the code fails to remove the listeners when the view is closed. Each event listener leaks `this`, which points to an instance of `Preview`.

### 2.1. Leak debugging via heap snapshots

There are currently no automated techniques for identifying memory leaks in web applications. The current state of the art is manual processing of heap snapshots. As we show, this approach does not effectively identify leaking objects or provide useful diagnostic information, and it thus does little to help developers locate and fix memory leaks.

The most popular way to manually debug memory leaks is via the *three heap snapshot technique* introduced by the Gmail team.<sup>8</sup> Developers repeat a task twice on a webpage and examine still-live objects created from the first run of the task. The assumption is that each run will clear out most of the objects created from the previous run and leave behind only leaking objects; in practice, it does not.

To apply this technique to Firefox’s debugger, the developer takes a heap snapshot after loading the debugger, a second snapshot after opening a source file, and a third snapshot after closing and reopening a source file. Then, the developer filters the third heap snapshot to focus only on objects allocated between the first and second.

This filtered view, as shown in Figure 2a, does not clearly identify a memory leak. Most of these objects are simply reused from the previous execution of the task and are not actually leaks, but developers must manually inspect these objects before they can come to that conclusion. The top item, `Array`, conflates all arrays in the application under one heading because JavaScript is dynamically typed. Confusingly, the entry (`array`) just below it refers to internal V8 arrays, which are not under the application’s direct control. Developers would be unlikely to suspect the `Preview` object, the primary leak, because it both appears low on the list and has a small retained size.

Even if a developer identifies a leaking object in a snapshot, it remains challenging to diagnose and fix because the snapshot contains no relation to code. The snapshot only provides retaining paths in the heap, which are often controlled by a third-party library or the browser itself. As Figure 2b shows, the retaining paths for a leaking `Preview` object



**Figure 2. The manual memory leak debugging process:** Currently, developers debug leaks by first examining heap snapshots to find leaking objects (a). Then, they try to use retaining paths to locate the code responsible (b). Unfortunately, these paths have no connection to code, so developers must search their codebase for identifiers referenced in the paths (see Section 2.1). This process can be time-consuming and ultimately fruitless. BLEAK saves considerable developer effort by automatically detecting and locating the code responsible for memory leaks.

Constructor	Distance	Objects Count	Shallow Size	Retained Size
▶ Array	4	3 143 0 %	100 576 0 %	31 099 584 26 %
▶ (array)	4	8 190 0 %	24 387 568 20 %	24 497 176 21 %
▶ BranchChunk	5	592 0 %	33 152 0 %	7 496 720 6 %
▶ LeafChunk	5	2 382 0 %	114 336 0 %	7 385 168 6 %
▶ Line	4	59 549 4 %	4 287 528 4 %	6 717 288 6 %
▶ (string)	5	3 823 0 %	4 761 800 4 %	4 761 800 4 %
▶ (sliced string)	5	55 931 4 %	2 237 240 2 %	2 237 240 2 %
▶ Doc	3	1 0 %	200 0 %	1 965 840 2 %
▶ Preview	6	1 0 %	200 0 %	489 016 0 %

(a) A truncated heap snapshot of the Firefox debugger, filtered using the three snapshot technique. The only relevant item is `Preview`, which appears low on the list underneath nonleaking objects.

Object	Distance	Shallow Size	Retained Size
▶ bound_this in native_bind() @4881	5	48 0 %	48 0 %
▶ [6] in Array @788047	4	32 0 %	64 0 %
▶ 0 in (object elements)[] @30597	5	32 0 %	32 0 %
▶ onScroll in Preview @488119	6	200 0 %	489 016 0 %
▶ this in system / Context @397635	5	56 0 %	56 0 %
▶ context in {} @387967	4	72 0 %	160 0 %
▶ native in HTMLDivElement @362	3	40 0 %	400 0 %
▶ [97] in Document DOM tree /	2	0 0 %	0 0 %
▶ 1 in (Document DOM trees)	1	0 0 %	0 0 %

(b) The retaining paths for `Preview`, the primary leaking object in the Firefox debugger. Finding the code responsible for leaking this object involves searching the entire production code base for identifiers in the retaining paths, which are commonly managed by third-party libraries and obfuscated via minification.

stem from an array and an unidentified DOM object. Locating the code responsible for a leak using these retaining paths involves grepping through the code for instances of the identifiers along the path. This task is often further complicated by two factors: (1) the presence of third-party libraries, which must be manually inspected; and (2) the common use of minification, which effectively obfuscates code and heap paths by reducing most variable names and some object properties to single letters.

### 3. BLEAK OVERVIEW

This section presents an overview of the techniques BLEAK uses to automatically detect, rank, and diagnose memory leaks. We illustrate these by showing how to use BLEAK to debug the Firefox memory leak presented in Section 2.

**Input script:** Developers provide BLEAK with a simple script that drives a web application in a loop through specific visual states. A *visual state* is the resting state of the GUI after the user takes an action, such as clicking on a link or submitting a form. The developer specifies the loop as an array of objects, where each object represents a specific visual state, comprising (1) a *check* function that checks the preconditions for being in that state, and (2) a *transition* function next that interacts with the page to navigate to the next visual state in the loop. The final visual state in the loop array transitions back to the first, forming a loop.

Figure 3a presents a loop for the Firefox debugger that opens and closes a source file in the debugger’s text editor. The first visual state occurs when there are no tabs open in the editor (line 8), and the application has loaded the list of documents in the application it is debugging (line 10); this is the default state of the debugger when it first loads. Once the application is in that first visual state, the loop transitions the application to the second visual state by clicking on `main.js` in the list of documents to open it in the text editor (line 12). The application reaches the second visible state once the debugger displays the contents of `main.js`

(line 18). The loop then closes the tab containing `main.js` (line 24), transitioning back to the first visual state.

**Locating leaks:** From this point, BLEAK proceeds entirely automatically. BLEAK uses the developer-provided script to drive the web application in a loop. Because object instances can change from snapshot to snapshot, BLEAK tracks *paths* instead of objects, letting it spot leaks even when a variable or object property is regularly updated with a new and larger object. For example, `history = history.concat(newItems)` overwrites `history` with a new and larger array.

During each visit to the first visual state in the loop, BLEAK takes a heap snapshot and tracks specific paths from GC roots that are continually growing. BLEAK treats a path as growing if the object identified by that path gains more outgoing references (e.g., when an array expands or when properties are added to an object).

For the Firefox debugger, BLEAK notices four heap paths that are growing each round trip: (1) an array within the `codeMirror` object that contains `scroll` event listeners, and internal browser event listener lists for (2) `mouseover`, (3) `mouseup`, and (4) `mousedown` events on the DOM element containing the text editor. Since these objects continue to grow over multiple loop iterations (the default setting is eight), BLEAK marks these items as *leak roots* as they appear to be growing without bound.

**Ranking leaks:** BLEAK uses the final heap snapshot and the list of leak roots to rank leaks by return on investment using a novel but intuitive metric we call *LeakShare* (Section 4.3) that prioritizes memory leaks that free the most memory with the least effort. *LeakShare* prunes objects in the graph reachable by nonleak roots and then splits the credit for remaining objects equally among the leak roots that retain them. Unlike retained size (a standard metric used by all existing heap snapshot tools), which only considers objects *uniquely owned* by leak roots, *LeakShare* correctly distributes the credit for the leaked `Preview` objects among the four different leak roots since they *all* must be removed to eliminate the leak.

**Figure 3. Automatic memory leak debugging with BLEAK: The only input developers need to provide to BLEAK is a simple script that drives the target web application in a loop (a). BLEAK then runs automatically, producing a ranked list of memory leaks with stack traces pointing to the code responsible for the leaks (b).**

```

1  exports.loop = [// Repeatedly open and close a source document.
2  { // Open a source document in the text editor. # Leak Root 1 [LeakShare: 811920]
3    check: function() {
4      const nodes = $('<code>.node</code>'); ## Leak Paths
5      // No documents are open
6      return $('<code>.source-tab</code>').length === 0 && * Event listeners for 'mouseover' on window.cm.display.wrapper
7      // Target document appears in doc list
8      nodes.length > 1 && nodes[1].innerText === "main.js"; ## Stack Traces Responsible
9    },
10   next: function() { $('<code>.node</code>')[1].click(); }
11 }, { // Close the document after it loads.
12   check: function() {
13     // Contents of main.js are in editor
14     return $('<code>.CodeMirror-line</code>').length > 2 &&
15     // Editor displays a tab for main.js
16     $('<code>.source-tab</code>').length === 1 &&
17     // Tab contains a close button
18     $('<code>.close-btn</code>').length === 1;
19   },
20   next: function() { $('<code>.close-btn</code>').click(); }
21 }];

```

(a) This script runs the Firefox debugger in a loop and is the only input BLEAK requires to automatically locate memory leaks. For brevity, we modify the script to use jQuery syntax.

906358432530

(b) A snippet from BLEAK's memory leak report for the Firefox debugger. BLEAK points directly to the code in Figure 1 responsible for the memory leak.

**Diagnosing leaks:** BLEAK next reloads the application and uses its proxy to transparently rewrite all of the JavaScript on the page, exposing otherwise-hidden edges in the heap as object properties. BLEAK uses JavaScript reflection to instrument identified leak roots to capture stack traces when they grow and when they are overwritten (not just where they were allocated). With this instrumentation in place, BLEAK uses the developer-provided script to run one final iteration of the loop to collect stack traces. These stack traces directly zero in on the code responsible for leak growth.

**Output:** Finally, BLEAK outputs its diagnostic report: a ranked list of leak roots (ordered by LeakShare), together with the heap paths that retain them and stack traces responsible for their growth. Figure 3b displays a snippet from BLEAK's output for the Firefox debugger, which points directly to the code responsible for the memory leak from Figure 1. With this information in hand, we were able to quickly develop a fix that removes the event listeners when the user closes the document. This fix has been incorporated into the latest version of the debugger.

## 4. ALGORITHMS

This section formally describes the operation of BLEAK's core algorithms for detecting (Section 4.1), diagnosing (Section 4.2), and ranking leaks (Section 4.3).

### 4.1. Memory leak detection

The input to BLEAK's memory leak detection algorithm is a set of heap snapshots collected during the same visual state, and the output is a set of *paths* from GC roots that are growing across all snapshots. We call these paths *leak roots*. BLEAK considers a path to be *growing* if the object at that path has more outgoing references than it did in the previous snapshot. To make the algorithm tractable, BLEAK only considers the shortest path to each specific heap item.

**Figure 4. PROPAGATEGROWTH propagates a node's growth status (*n.growing*) between heap snapshots. BLEAK considers a path in the heap to be growing if the node at the path continually increases its number of outgoing edges.**

```

PROPAGATEGROWTH( $G, G'$ )
1   $Q = [(G.root, G'.root)]$ ,  $G'.root.mark = \text{TRUE}$ 
2  for each node  $n \in G'.N$ 
3     $n.growing = \text{FALSE}$ 
4  while  $|Q| > 0$ 
5     $(n, n') = \text{DEQUEUE}(Q)$ 
6     $E_n = \text{GETOUTGOINGEDGES}(G, n)$ 
7     $E'_n = \text{GETOUTGOINGEDGES}(G', n')$ 
8     $n'.growing = n.growing \wedge |E_n| < |E'_n|$ 
9    for each edge  $(n_1, n_2, l) \in E_n$ 
10     for each edge  $(n'_1, n'_2, l') \in E'_n$ 
11       if  $l == l'$  and  $n'_2.mark == \text{FALSE}$ 
12          $n'_2.mark = \text{TRUE}$ 
13          $\text{ENQUEUE}((n_2, n'_2))$ 

```

Each heap snapshot contains a heap graph  $G = (N, E)$  with a set of nodes  $N$  that represent items in the heap and edges  $E$  where each edge  $(n_1, n_2, l) \in E$  represents a reference from node  $n_1$  to  $n_2$  with label  $l$ . A label  $l$  is a tuple containing the type and name of the edge. Each edge's type is either a *closure variable* or an *object property*. An edge's name corresponds to the name of the closure variable or object property. For example, the object  $O = \{\text{foo}: 3\}$  has an edge  $e$  from  $O$  to the number 3 with label  $l = (\text{property}, \text{"foo"})$ . A path  $P$  is simply a list of edges  $(e_1, e_2, \dots, e_n)$  where  $e_1$  is an edge from the root node  $(G.root)$ .<sup>2</sup>

For the first heap snapshot, BLEAK conservatively marks every node as *growing*. For subsequent snapshots, BLEAK runs PROPAGATEGROWTH (Figure 4) to propagate the growth flags from the previous snapshot to the new snapshot and discards the previous snapshot. On line 2, PROPAGATEGROWTH

<sup>2</sup> For simplicity, we describe heap graphs as having one root.



initializes every node in the new graph to *not growing* to prevent spuriously marking new growth as growing in the next run of the algorithm. Since the algorithm only considers paths that are the shortest path to a specific node, it is able to associate growth information with the terminal node, which represents a specific path in the heap.

PROPAGATEGROWTH runs a breadth-first traversal across shared paths in the two graphs, starting from the root node that contains the global scope (`window`) and the DOM. The algorithm marks a node in the new graph as *growing* if the node at the same path in the previous graph is both growing and has fewer outgoing edges (line 8). As a result, the algorithm will only mark a heap path as a leak root if it consistently grows between every snapshot and if it has been present since the first snapshot.

PROPAGATEGROWTH only visits paths shared between the two graphs (line 11). At a given path, the algorithm considers an outgoing edge  $e_n$  in the old graph and  $e'_n$  in the new graph as equivalent if they have the same label. In other words, the edges have to correspond to the same property name on the object at that path, or a closure variable with the same name captured by the function at that path.

After propagating growth flags to the final heap snapshot, BLEAK runs FINDLEAKPATHS (Figure 5) to record growing paths in the heap. This traversal visits *edges* in the graph to capture the shortest path to all unique edges that point to growing nodes. For example, if a growing object  $O$  is located at `window.O` and as variable  $p$  in the function `window.L.z`, FINDLEAKPATHS will report both paths. This property is important for diagnosing leaks, as we discuss in Section 4.2.

BLEAK takes the output of FINDLEAKPATHS and groups it by the terminal node of each path. Each group corresponds to a specific leak root. This set of leak roots forms the input to the ranking algorithm.

## 4.2. Diagnosing leaks

Given a list of leak roots and, for each root, a list of heap paths that point to the root, BLEAK diagnoses leaks through hooks that run whenever the application performs any of the following actions:

**Figure 5. FINDLEAKPATHS, which returns paths through the heap to leaking nodes. The algorithm encodes each path as a list of edges formed by tuples (t).**

```

FINDLEAKPATHS( $G$ )
1   $Q = []$ ,  $T_{Gr} = \{\}$ 
2  for each edge  $e = (n_1, n_2, l) \in G.E$  where  $n_1 == G.root$ 
3       $e.mark = \text{TRUE}$ 
4      ENQUEUE( $Q$ , (NIL,  $e$ ))
5  while  $|Q| > 0$ 
6       $t = \text{DEQUEUE}(Q)$ 
7       $(t_p, (n_1, n_2, l)) = t$ 
8      if  $n_2.growing == \text{TRUE}$ 
9           $T_{Gr} = T_{Gr} \cup \{t\}$ 
10     for each edge  $e = (n'_1, n'_2, l') \in G.E$ 
11         if  $n'_1 == n_2$  and  $e.mark == \text{FALSE}$ 
12              $e.mark = \text{TRUE}$ 
13             ENQUEUE( $Q$ , ( $t, e$ ))
14  return  $T_{Gr}$ 

```

- *Grows a leak root* with a new item. This growth occurs when the application adds a property to an object, an element to an array, an event listener to an event target, or a child node to a DOM node. BLEAK captures a stack trace and associates it with the new item.
- *Shrinks a leak root* by removing any of the previously-mentioned items. BLEAK removes any stack traces associated with the removed items, as the items are no longer contributing to the leak root's growth.
- *Assigns a new value to a leak root*, which typically occurs when the application copies the state from an old version of the leaking object into a new version. BLEAK removes all previously-collected stack traces for the leak root, collects a new stack trace, associates it with all of the items in the new value, and inserts the grow and shrink hooks into the new value.

BLEAK runs one loop iteration of the application with all hooks installed. This process generates a list of stack traces responsible for growing each leak root.

## 4.3. Leak root ranking

BLEAK uses a new metric to rank leak roots by return on investment that we call *LeakShare*. LeakShare prioritizes memory leaks that free the most memory with the least effort by dividing the "credit" for retaining a shared leaked object equally among the leak roots that retain them.

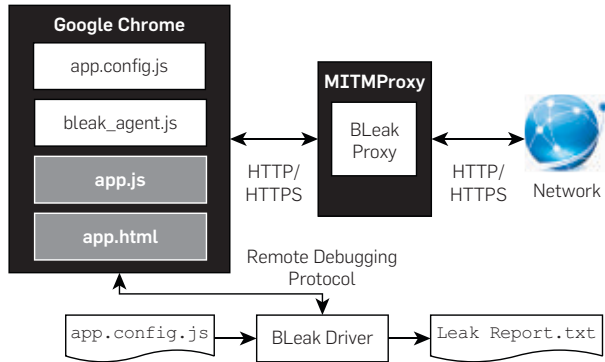
LeakShare first marks all of the items in the heap that are reachable from nonleaks via a breadth-first traversal that stops at leak roots. These nodes are ignored by subsequent traversals. Then, LeakShare performs a breadth-first traversal from each leak root that increments a counter on all reachable nodes. Once this process is complete, every node has a counter containing the number of leak roots that can reach it. Finally, the algorithm calculates the LeakShare of each leak root by adding up the size of each reachable node divided by its counter, which splits the "credit" for the node among all leak roots that can reach it. Our PLDI paper presents the full algorithm for LeakShare.<sup>15</sup>

## 5. IMPLEMENTATION

BLEAK consists of three main components that work together to automatically debug memory leaks (see Figure 6): (1) a driver program orchestrates the leak debugging process; (2) a proxy transparently performs code rewriting on-the-fly on the target web application; and (3) an agent script embedded within the application exposes hidden state for leak detection and growth events for leak diagnosis. We briefly describe how these components work here; our PLDI paper provides further details.<sup>15</sup>

To initiate leak debugging, the BLEAK driver launches BLEAK's proxy and the Google Chrome browser with an empty cache, a fresh user profile, and a configuration that uses the BLEAK proxy. The driver connects to the browser via the standard Chrome DevTools Protocol, navigates to the target web application, and uses the developer-provided configuration file to drive the application in a loop. During each repeat visit to the first visual state in the loop, the driver takes a heap snapshot via the remote debugging protocol

**Figure 6. BLEAK system overview. White items are BLEAK components, gray items are rewritten by the proxy during leak diagnosis, and black items are unmodified.**



and runs PROPAGATEGROWTH (Figure 4) to propagate growth information between heap snapshots.

At the end of a configurable number of loop iterations (the default is 8), the driver shifts into diagnostic mode. The driver runs FINDLEAKPATHS to locate all of the paths to all of the leak roots (Figure 5), configures the proxy to perform code rewriting for diagnosis, and reloads the page to pull in the transformed version of the web application. The driver runs the application in a single loop iteration before triggering the BLEAK agent to insert diagnostic hooks that collect stack traces at all of the paths reported by FINDLEAKPATHS. Then, the driver runs the application in a final loop before retrieving stack traces from the agent. Finally, the driver runs LeakShare (Section 4.3) to rank leak roots and generate a memory leak report.

## 6. EVALUATION

We evaluate BLEAK by running it on production web applications. Our evaluation addresses the following questions:

- **Precision:** How precise is BLEAK’s memory leak detection? (Section 6.2)
- **Accuracy of diagnoses:** Does BLEAK accurately locate the code responsible for memory leaks? (Section 6.2)
- **Impact of discovered leaks:** How impactful are the memory leaks that BLEAK finds? (Section 6.3)
- **Utility of ranking:** Is LeakShare an effective metric for ranking the severity of memory leaks? (Section 6.4)

Our evaluation finds **59 distinct memory leaks** across five web applications, *all of which were unknown to application developers*. Of these, 27 corresponded to known-but-unfixed memory leaks in JavaScript library dependencies, of which only 6 were independently diagnosed and had pending fixes. We reported all 32 new memory leaks to the relevant developers along with our fixes; 16 are now fixed, and 4 have fixes in code review. We find new leaks in popular applications and libraries including Airbnb, Angular JS (1.x), Google Maps SDK, Google Tag Manager, and Google Analytics.

We run BLEAK on each web application for 8 round trips through specific visual states to produce a BLEAK leak report, as shown in Figure 3b. We describe these loops using only 17–73 LOC. This process takes less than 15 min per

application on our evaluation machine, a MacBook Pro with a 2.9GHz Intel Core i5 and 16GB of RAM. For each application, we analyze the reported leaks, write a fix for each true leak, measure the impact of fixing the leaks, and compare LeakShare with alternative ranking metrics.

### 6.1. Applications

Because there is no existing corpus of benchmarks for web application memory leak detection, we created one. Our corpus consists of five popular web applications that both comprise large code bases and whose overall memory usage appeared to be growing over time. We primarily focus on open source web applications because it is easier to develop fixes for the original source code; this represents the normal use case for developers. We also include a single closed-source website, Airbnb, to demonstrate BLEAK’s ability to diagnose websites in production. We present each web application, highlight a selection of the libraries they use, and describe the loop of visual states we use in our evaluation:

**Airbnb:** A website offering short-term rentals and other services, Airbnb uses React, Google Maps SDK, Google Analytics, the Criteo OneTag Loader, and Google Tag Manager. BLEAK loops between the pages `/s/all`, which lists all services offered on Airbnb, and `/s/homes`, which lists only homes and rooms for rent.

**Piwik 3.0.2:** A widely-used open-source analytics platform; we run BLEAK on its in-browser dashboard that displays analytics results. The dashboard primarily uses jQuery and AngularJS. BLEAK repeatedly visits the main dashboard page, which displays a grid of widgets.

**Loomio 1.8.66:** An open-source collaborative platform for group decision-making. Loomio uses AngularJS, LokiJS, and Google Tag Manager. BLEAK runs Loomio in a loop between a group page, which lists all of the threads in that group, and the first thread listed on that page.

**Mailpile v1.0.0:** An open-source mail client. Mailpile uses jQuery. BLEAK runs Mailpile’s demo in a loop that visits the inbox and the first four emails in the inbox.

**Firefox Debugger (commit 91f5c63):** An open-source JavaScript debugger written in React that runs in any web browser. We run the debugger while it is attached to a Firefox instance running Mozilla’s SensorWeb. BLEAK runs the debugger in a loop that opens and closes SensorWeb’s `main.js` in the debugger’s text editor.

### 6.2. Precision and accuracy

To determine BLEAK’s leak detection precision and the accuracy of its diagnoses, we manually check each BLEAK-reported leak in the final report to confirm (1) that it is growing without bound and (2) that the stack traces correctly report the code responsible for the growth. Figure 8 summarizes our results.

**Bleak has an average precision of 96.8% and a median precision of 100%** on our evaluation applications. There



are only three false positives. All point to an object that continuously grows until some threshold or timeout occurs; developers using BLEAK can avoid these false positives by increasing the number of round trips. Two of the three false positives are actually the same object located in the Google Tag Manager JavaScript library.

**With one exception, BLEAK accurately identifies the code responsible for all of the true leaks.** BLEAK reports stack traces that directly identify the code responsible for each leak. In cases where multiple independent source locations grow the same leak root, BLEAK reports all relevant source locations. For one specific memory leak, BLEAK fails to record a stack trace. **Guided by BLEAK's leak reports, we were able to fix every memory leak.** Each memory leak took approximately 15 min to fix.

### 6.3. Leak impact

To determine the impact of the memory leaks that BLEAK reports, we measure each application's live heap size over 10 loop iterations with and without our fixes. We use BLEAK's HTTP/HTTPS proxy to directly inject memory leak fixes into the application, which lets us test fixes on closed-source websites like Airbnb. We run each application except Airbnb 5 times in each configuration (we run Airbnb only once per configuration for reasons discussed in Section 6.4).

To calculate the leaks' combined impact on overall heap growth, we calculate the average live heap growth between loop iterations with and without the fixes in place and take the difference (Growth Reduction). For this metric, we ignore the first five loop iterations because these are noisy due to application startup. Figures 7 and 8 present the results.

**On average, fixing the memory leaks that BLEAK reports eliminates over 93% of all heap growth on our benchmarks (median: 98.2%).** These results suggest that BLEAK does not miss any significantly impactful leaks.

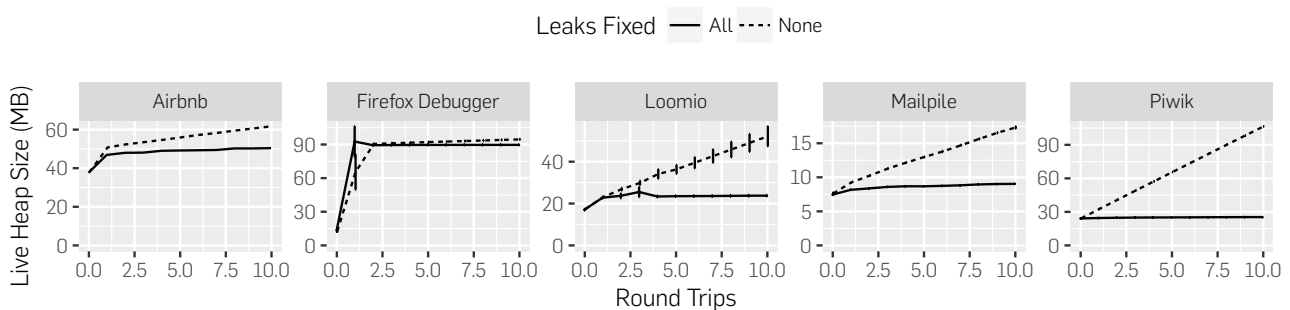
### 6.4. LeakShare effectiveness

We compare LeakShare against two alternative ranking metrics: retained size and transitive closure size. Retained size corresponds to the amount of memory the garbage collector would reclaim if the leak root were removed from the heap graph and is the metric that standard heap snapshot viewers display to the developer. The transitive closure size of a leak root is the size of all objects reachable from the leak root as used by Xu et al.<sup>16</sup> Since JavaScript heaps are highly connected and frequently contain references to the global scope, we expect this metric to report similar values for most leaks.

We measure the effectiveness of each ranking metric by calculating the growth reduction (as in Section 6.3) over the application with no fixes after fixing each memory leak in ranked order. We then calculate the quartiles of this data, indicating how much heap growth is eliminated after fixing the top 25%, 50%, and 75% of memory leaks reported ranked by a given metric. We sought to write patches for each evaluation application that fix a single leak root at a time, but this is not feasible in all cases; some leaks share the same root cause. In these cases, we apply the patch during a ranking for the first relevant leak root reported.

We run each application except Airbnb for ten loop iterations over five runs for each unique combination of metric and number of top-ranked leak roots to fix. We avoid running duplicate configurations when multiple metrics report the same ranking. Airbnb is challenging to evaluate

**Figure 7. Impact of fixing memory leaks found with BLEAK: Graphs display live heap size over round trips; error bars indicate the 95% confidence interval. Fixing the reported leaks eliminates an average of 93% of all heap growth.**



**Figure 8. BLEAK precisely finds impactful memory leaks: On average, BLEAK finds these leaks with over 95% precision, and fixing them eliminates over 90% of all heap growth.**

Program	Loop LOC	Leak Roots	False Positives	Distinct Leaks	Precision	Growth Reduction
Airbnb	17	32	2	32	94%	1.04 MB (81.0%)
Piwik	32	17	0	11	100%	8.14 MB (99.3%)
Loomio	73	10	1	9	90%	2.83 MB (98.3%)
Mailpile	37	4	0	3	100%	0.80 MB (91.8%)
Firefox Debugger	17	4	0	4	100%	0.47 MB (98.2%)
<b>Total / mean:</b>	35	67	3	59	96.8%	2.66 MB (93.7%)

**Figure 9. Performance of ranking metrics: Growth reduction by metric after fixing quartiles of top-ranked leaks. Bold indicates greatest reduction ( $\pm 1\%$ ). We omit Firefox; it has only four leaks, which must all be fixed (see Section 2). LeakShare generally outperforms or matches other metrics.**

Growth Reduction for Top Leaks Fixed				
Program	Metric	25%	50%	75%
Airbnb	LeakShare	OK	111K	<b>462K</b>
	Retained Size	OK	OK	105K
	Trans. Closure Size	OK	<b>196K</b>	393K
Loomio	LeakShare	OK	<b>1083K</b>	<b>2878K</b>
	Retained Size	<b>64K</b>	186K	<b>2898K</b>
	Trans. Closure Size	59K	67K	2398K
Mailpile	LeakShare	<b>613K</b>	<b>817K</b>	<b>820K</b>
	Retained Size	<b>613K</b>	<b>817K</b>	<b>820K</b>
	Trans. Closure Size	OK	OK	201K
Piwik	LeakShare	<b>8003K</b>	<b>8104K</b>	<b>8306K</b>
	Retained Size	2073K	7969K	<b>8235K</b>
	Trans. Closure Size	103K	110K	374K

because it has 30 leak roots, randomly performs A/B tests between runs, and periodically updates its minified codebase in ways that break our memory leak fixes. As a result, we were only able to gather one run of data for Airbnb for each unique configuration. Figure 9 displays the results.

**In most cases, LeakShare outperforms or ties the other metrics.** LeakShare initially is outperformed by other metrics on Airbnb and Loomio because it prioritizes leak roots that share significant state with other leak roots. Retained size always prioritizes leak roots that uniquely own the most state, which provide the most growth reduction in the short term. LeakShare eventually surpasses the other metrics on these two applications as it fixes the final leak roots holding on to shared state.

## 7. RELATED WORK

**Web application memory leak detectors:** BLEAK automatically debugs memory leaks in web applications; past work in this space is ineffective or not sufficiently general. LeakSpot locates allocation and reference sites that produce and retain increasing numbers of objects over time and uses staleness as a heuristic to refine its output.<sup>14</sup> On real applications, LeakSpot typically reports over 50 different allocation and reference sites that developers must manually inspect to identify and diagnose memory leaks. JSWhiz statically analyzes code written with Google Closure type annotations to detect specific leak patterns.<sup>13</sup>

**Web application memory debugging:** Some tools help web developers debug memory usage and present diagnostic information that developers must manually interpret to locate leaks (Section 2 describes Google Chrome's DevTools). MemInsight summarizes and displays information about the JavaScript heap, including per-object-type staleness information, the allocation site of objects, and retaining paths in the heap.<sup>7</sup> Unlike BLEAK, these tools do not directly identify memory as leaking or identify the code responsible for leaks.

**Growth-based memory leak detection:** LeakBot looks for patterns in the heap graphs of Java applications to find

memory leaks.<sup>9</sup> LeakBot assumes that leak roots own all of their leaking objects, but leaked objects in web applications frequently have multiple owners. BLEAK does not rely on specific patterns and uses round trips to the same visual state to identify leaking objects.

**Staleness-based memory leak detection:** SWAT (C/C++), Sleigh (JVM), and Hound (C/C++) find leaking objects using a staleness metric derived from the last time an object was accessed and identify the call site responsible for allocating them.<sup>6, 2, 12</sup> Leakpoint (C/C++) also identifies the last point in the execution that referenced a leaking memory location.<sup>3</sup> Xu et al. identify leaks stemming from Java collections using a hybrid approach that targets containers that grow in size over time and contain stale items. As we discuss in our PLDI paper, staleness is ineffective for at least 77% of the memory leaks BLEAK identifies.<sup>15</sup>

## 8. CONCLUSION

This paper presents BLEAK, the first effective system for debugging client-side memory leaks in web applications. We show that BLEAK has high precision and finds numerous previously-unknown memory leaks in web applications and libraries. BLEAK is open source and is available for download at <http://bleak-detector.org/>.

## Acknowledgments

John Vilck was supported by a Facebook PhD Fellowship. This material is based upon work supported by the National Science Foundation under Grant No. 1637536. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. □

## References

- Basques, K. Fix memory problems, 2017. <https://developers.google.com/web/tools/chrome-devtools/memory-problems/>.
- Bond, M.D., McKinley, K.S. Bell: Bit-encoding online memory leak detection. In *ASPLOS*, ACM, San Jose, CA, 2006, 61–72.
- Clause, J.A., Orso, A. LEAKPOINT: Pinpointing the causes of memory leaks. In *ICSE*, ACM, Cape Town, South Africa, 2010, 515–524.
- Google. Speed up google chrome, 2017. <https://support.google.com/chrome/answer/1385029>.
- Hara, K. Oilpan: GC for blink, 2013. <https://docs.google.com/presentation/d/1YtufurcyKFS0hxP0nC3U6JJroM8aRP49Yf0QWznZ9jrk>.
- Hauswirth, M., Chilimbi, T.M. Low-overhead memory leak detection using adaptive statistical profiling. In *ASPLOS*, ACM, Boston, MA, 2004, 156–164.
- Jensen, S.H., Sridharan, M., Sen, K., Chandra, S. MemInsight: Platform-independent memory debugging for JavaScript. In *FSE*, ACM, Bergamo, Italy, 2015, 345–356.
- Lee, L., Hundt, R. BloatBusters: Eliminating memory leaks in Gmail, 2012. <https://docs.google.com/presentation/d/1wUVmf78gG-ra5aOxvTfYdiLkdGaR9OhXRnOLICEmu2s>.
- Mitchell, N., Sevitsky, G. LeakBot: An automated and lightweight tool for diagnosing memory leaks in large Java applications. In *ECOOP*, 2003, 351–377.
- Mozilla. about:memory, 2017. <https://developer.mozilla.org/en-US/docs/Mozilla/Performance/about:memory>.
- Nguyen, N. The best firefox ever, 2017. <https://blog.mozilla.org/blog/2017/06/13/faster-better-firefox/>.
- Novark, G., Berger, E.D., Zorn, B.G. Efficiently and precisely locating memory leaks and bloat. In *PLDI*, ACM, Dublin, Ireland, 2009, 397–407.
- Pienaar, J.A., Hundt, R. JSWhiz: Static analysis for JavaScript memory leaks. In *CGO*, IEEE Computer Society, Shenzhen, China, 2013, 11:1–11:11.
- Rudafshani, M., Ward, P.A.S. Leakspot: Detection and diagnosis of memory leaks in javascript applications. *Softw. Pract. Exp.* 1, 47 (2017), 97–123.
- Vilk, J., Berger, E.D. BLEAK: Automatically debugging memory leaks in web applications. In *PLDI*, ACM, Philadelphia, PA, 2018, 15–29.
- Xu, G.H., Rountev, A. Precise memory leak detection for Java software using container profiling. *TOSEM* 3, 22 (2013):17:1–17:28.

John Vilck and Emery D. Berger ({jvilk, emery}@cs.umass.edu), College of Information and Computer Sciences, University of Massachusetts Amherst.



volume

01

number

01

FIRST

ISSUE

PUBLISHED

*Digital Threats:  
Research and Practice*  
is now available in  
the ACM Digital Library



*Digital Threats: Research and Practice* (DTRAP) is a peer-reviewed open access 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.



Association for  
Computing Machinery

<https://dtrap.acm.org>

## Boston College

### Non Tenure-Track Position in Computer Science

The Computer Science Department of Boston College seeks to fill one or possibly more non-tenure track teaching positions, as well as shorter-term visiting teaching positions. **One of these positions has a January, 2021 start date.** All applicants should be committed to excellence in undergraduate education and be able to teach a broad variety of undergraduate computer science courses. We are especially interested in candidates who are able to teach courses in systems and networks. Faculty in longer-term positions will also participate in the development of new courses that reflect the evolving landscape of the discipline.

Minimum requirements for the title of Assistant Professor of the Practice, and for the title of Visiting Assistant Professor, include a Ph.D. in Computer Science or closely related discipline.

Candidates without a Ph.D. would be eligible for the title of Lecturer or Visiting Lecturer.

We will begin reviewing applications as they are received and will continue considering applications until the positions are filled. Applicants should submit a cover letter, CV, and a separate teaching statement and arrange for three confidential letters of recommendation that comment on their teaching performance to be uploaded directly to Interfolio. To apply go to: <http://apply.interfolio.com/78108>

Boston College conducts background checks as part of the hiring process. Information about the University and our department is available at [bc.edu](http://bc.edu) and [cs.bc.edu](http://cs.bc.edu).

Boston College is a Jesuit, Catholic university that strives to integrate research excellence with a foundational commitment to formative liberal arts education. We encourage applications from candidates who are committed to fostering a diverse and inclusive academic community. Boston College is an Affirmative Action/Equal Opportunity Employer and does not discriminate on the basis of any legally protected category including disability and protected veteran status. To learn more about how BC supports diversity and inclusion throughout the university, please visit the Office for Institutional Diversity at <http://www.bc.edu/offices/diversity>.

## California Institute of Technology

### Faculty Position in Computing and Mathematical Sciences

The Computing and Mathematical Sciences (CMS) Department at the California Institute of Technology (Caltech) invites applications for tenure-track faculty positions. The CMS Department is part of the Division of Engineering and Applied Science (EAS), comprising researchers working in and between the fields of aerospace, civil, electrical, environmental, mechanical, and medical

engineering, as well as materials science and applied physics. The Institute as a whole represents the full range of research in biology, chemistry, engineering, geological and planetary sciences, physics, and the social sciences.

Fundamental research in computing and mathematical sciences, and applied research which links to activities in other parts of Caltech, are both welcomed. A commitment to world-class research, as well as high-quality teaching and mentoring, is expected, and appointment as an assistant professor is contingent upon the completion of a Ph.D. degree in applied mathematics, computer science or related areas. The initial appointment at the assistant professor level is four years. Reappointment beyond the initial term is contingent upon successful review conducted prior to the commencement of the fourth year.

► Interviews will take place in January and February 2021.

► Applications will be reviewed beginning 22 October 2020 and all applications received before 1 December 2020 will receive full consideration.

► Applications received before 8 November will be considered for interviews in January.

► Applications received after 8 November will be considered for interviews in February.

To fulfill Caltech's commitment to promoting diversity, inclusiveness, and excellence in research on our campus, we actively seek candidates who can work with, teach, and mentor students from under-represented communities. Along with other standard application materials, applicants should submit a diversity and inclusion statement that discusses past and/or anticipated contributions to improving diversity, equity, and inclusion in the areas of research, teaching, and/or outreach.

For a list of all documents required, and full instructions on how to apply online, please visit <https://applications.caltech.edu/jobs/cms>. Questions about the application process may be directed to [search@cms.caltech.edu](mailto:search@cms.caltech.edu).

Caltech is an equal opportunity employer and all qualified applicants will receive consideration for employment without regard to age, race, color, religion, sex, sexual orientation, gender identity, national origin, disability status, protected veteran status, or any other characteristic protected by law.



## Faculty Positions in Computer Science

The Department of Computer Science at the National University of Singapore (NUS) invites applications for tenure-track and educator-track positions in all areas of computer science.

The Department is looking for candidates for all levels of tenured and tenure-track positions in any area of computer science. Candidates for Assistant Professor positions on the tenure track should be early in their academic careers and yet demonstrate outstanding research potential, and a strong commitment to teaching.

For Senior Lecturer and Associate Professor on the educator-track, teaching experience or relevant industry experience will be preferred. Besides relevant background and experience, we are also looking for someone with a passion for imparting the latest knowledge in computing to students in our programs

The Department enjoys ample research funding, moderate teaching loads, excellent facilities, and extensive international collaborations. We have a full range of faculty covering all major research areas in computer science and boasts a thriving PhD program that attracts the brightest students from the region and beyond. More information is available <https://www.comp.nus.edu.sg/about/depts/cs/recruitment/faculty/>

NUS is an equal opportunity employer that offers highly competitive salaries, and is situated in Singapore, an English-speaking cosmopolitan city that is a melting pot of many cultures, both the east and the west. Singapore offers high-quality education and healthcare at all levels, as well as very low tax rates.

### Application Details:

- Submit the following documents (in a single PDF) online via: <https://faces.comp.nus.edu.sg>
  - A cover letter that indicates the position applied for and the main research interests
  - Curriculum Vitae
  - A teaching statement
  - A research statement (optional but encouraged for educator-track)
- Provide the contact information of 3 referees when submitting your online application, or, arrange for at least 3 references to be sent directly to [csrec@comp.nus.edu.sg](mailto:csrec@comp.nus.edu.sg)
- To ensure maximal consideration, please submit your application by 18 December 2020.
- If you have further enquiries, please contact the Search Committee Chair, Joxan Jaffar, at [csrec@comp.nus.edu.sg](mailto:csrec@comp.nus.edu.sg)



## Georgia Institute of Technology

### Tenure-Track Faculty

The School of Computational Science and Engineering (CSE) in the College of Computing at the Georgia Institute of Technology invites applications for multiple openings at the Assistant Professor level (tenure-track); exceptional candidates at the Associate Professor and Professor level also will be considered. CSE focuses on foundational research of an interdisciplinary nature that enables advances in science, engineering, medical, and social domains. Applicants are expected to develop and sustain a research program in one or more of our core areas: high-performance computing, scientific and numerical computing, modeling and simulation, discrete algorithms, and large-scale data analytics (including machine learning and artificial intelligence).

All areas of research will be considered, especially: scientific artificial intelligence (AI methods unique to scientific computing), urban computing (enabling effective design and operation of cities and urban communities), application-driven post-Moore's law computing, and data science for fighting disease. Applicants must have an outstanding record of research and a commitment to teaching.

Applicants are expected to engage in substantive research with collaborators in other disciplines. For example, current faculty have domain expertise and/or collaborations in computational chemistry; earth sciences; biomedical and health sciences; urban systems and smart cities; social good and sustainable development; materials and manufacturing; and others.

Georgia Tech is organized into six Colleges. The School of Computational Science and Engineering resides in the College of Computing along with the School of Computer Science and the School of Interactive Computing. Joint appointments with other Schools in the College of Computing as well as Schools in other Colleges will be considered.

Applications should be submitted online through: <https://academicjobsonline.org/ajo/jobs/16901>. The application materials should include a full academic CV, a personal narrative on teaching and research, at least three references, one sample publication that is considered a very significant research contribution, and the names of 2-3 CSE faculty members closest to the applicant's research (see <https://www.cse.gatech.edu/people/faculty> for current faculty). For full consideration, applications are due by December 1, 2020.

Georgia Tech is an Affirmative Action/Equal Opportunity Employer. Applications from women and under-represented minorities are strongly encouraged.

For more information about Georgia Tech's School of Computational Science and Engineering please visit: <http://www.cse.gatech.edu/>

## Indiana University

### Luddy School of Informatics, Computing, and Engineering

#### Assistant Professor in Computer Science

The Luddy School of Informatics, Computing, and Engineering at Indiana University (IU) Bloomington invites applications for a tenure track assistant

professor position in Computer Science to begin in Fall 2021. We are particularly interested in candidates with research interests in formal models of computation, algorithms, information theory, and machine learning with connection to quantum computation, quantum simulation, or quantum information science. The successful candidate will also be a Quantum Computing and Information Science Faculty Fellow supported in part for the first three years by an NSF-funded program that aims to grow academic research capacity in the computing and information science fields to support advances in quantum computing and/or communication over the long term. For additional information about the NSF award please visit: [https://www.nsf.gov/awardsearch/showAward?AWD\\_ID=1955027&HistoricalAwards=false](https://www.nsf.gov/awardsearch/showAward?AWD_ID=1955027&HistoricalAwards=false)

The position allows the faculty member to collaborate actively with colleagues from a variety of outside disciplines including the departments of physics, chemistry, mathematics and intelligent systems engineering, under the umbrella of the Indiana University funded "quantum science and engineering center" (IU-QSEC).

We seek candidates prepared to contribute to our commitment to diversity and inclusion in higher education, especially those with experience in teaching or working with diverse student populations. Duties will include research, teaching multi-level courses both online and in person, participating in course design and assessment, and service to the School. Applicants should have a demonstrable potential for excellence in research and teaching and a PhD in Computer Science or a related field expected before August 2021.

Candidates should review application requirements, learn more about the Luddy School and apply online at: <https://indiana.peopleadmin.com/postings/9841>.

For full consideration submit online application by December 1, 2020. Applications will be considered until the positions are filled. Questions may be sent to [sabry@indiana.edu](mailto:sabry@indiana.edu).

Indiana University is an equal employment and affirmative action employer and a provider of ADA services. All qualified applicants will receive consideration for employment without regard to age, ethnicity, color, race, religion, sex, sexual orientation, gender identity or expression, genetic information, marital status, national origin, disability status or protected veteran status.

## The Johns Hopkins University

### Lecturer/Sr. Lecturer in Computer Science

The Department of Computer Science at Johns Hopkins University seeks applicants for a full-time teaching position. This is a career-oriented, renewable appointment that is responsible for the development and delivery of undergraduate and graduate courses, depending on the candidate's background. These positions carry a 3 course load per semester, usually with only 2 different preps. Teaching faculty are also encouraged to engage in departmental and university service and may have advising responsibilities. Extensive grading support is given to all instructors. The university has instituted a non-tenure track career path for full-time teaching faculty culminating in the rank of Teaching Professor.

Johns Hopkins is a private university known for its commitment to academic excellence and



## TENURE-TRACK AND TENURED POSITIONS

### School of Information Science and Technology (SIST)

ShanghaiTech University invites highly qualified candidates to fill multiple tenure-track/tenured faculty positions as its core founding team in the School of Information Science and Technology (SIST). We seek candidates with exceptional academic records or demonstrated strong potentials in all cutting-edge research areas of information science and technology. They must be fluent in English. English-based overseas academic training or background is highly desired.

ShanghaiTech is founded as a world-class research university for training future generations of scientists, entrepreneurs, and technical leaders. Boasting a new modern campus in Zhangjiang Hightech Park of cosmopolitan Shanghai, ShanghaiTech shall trail-blaze a new education system in China. Besides establishing and maintaining a world-class research profile, faculty candidates are also expected to contribute substantially to both graduate and undergraduate educations.

**Academic Disciplines:** Candidates in all areas of information science and technology shall be considered. Our recruitment focus includes, but is not limited to: computer science and technology, electronic science and technology, information and communication engineering, applied mathematics and statistics, data science, robotics, bioinformatics, biomedical engineering, internet of things, smart energy, computer systems and security, operation research, mathematical optimization and other interdisciplinary fields involving information science and technology, especially areas related to AI.

**Compensation and Benefits:** Salary and startup funds are highly competitive, commensurate with experience and academic accomplishment. We also offer a comprehensive benefit package to employees and eligible dependents, including on-campus housing. All regular ShanghaiTech faculty members will join its new tenure-track system in accordance with international practice for progress evaluation and promotion.

#### Qualifications:

- Strong research productivity and demonstrated potentials;
- Ph.D. (Electrical Engineering, Computer Engineering, Computer Science, Statistics, Applied Math, or related field);
- A minimum relevant (including PhD) research experience of 4 years.

**Applications:** Submit (in English, PDF version) a cover letter, a 2-page research plan, a CV plus copies of 3 most significant publications, and names of three referees to: [sist@shanghaitech.edu.cn](mailto:sist@shanghaitech.edu.cn)

For more information, please visit: <http://sist.shanghaitech.edu.cn/>

**Deadline:** December 31, 2020

research. The Computer Science department is one of nine academic departments in the Whiting School of Engineering, on the beautiful Homewood Campus. We are located in Baltimore, MD in close proximity to Washington, DC and Philadelphia, PA. See the department webpage at <https://cs.jhu.edu> for additional information about the department, including undergraduate and graduate programs and current course descriptions.

Applicants for the position should have a Ph.D. in Computer Science or a closely related field. Demonstrated excellence in and commitment to teaching, and excellent communication skills are expected of all applicants. Applications may be submitted online at <http://apply.interfolio.com/78726>. Questions may be directed to [lecsearch2020@cs.jhu.edu](mailto:lecsearch2020@cs.jhu.edu). For full consideration, applications should be submitted by December 1, 2020. Applications will be accepted until the position is filled.

The Department is conducting a broad and inclusive search and is committed to identifying candidates who through their teaching and service will contribute to the diversity and excellence of the academic community.

The Johns Hopkins University is committed to active recruitment of a diverse faculty and student body. The University is an Affirmative Action/Equal Opportunity Employer of women, minorities, protected veterans and individuals with disabilities and encourages applications from these and other protected group members. Consistent with the University's goals of achieving excellence in all areas, we will assess the comprehensive qualifications of each applicant.

### **The Johns Hopkins University Tenure-Track Faculty, Department of Computer Science**

The Johns Hopkins University's Department of Computer Science seeks applicants for tenure-track faculty positions at all levels and across all areas of computer science. The department is particularly interested in applicants in the areas of computational biology, bioinformatics, human-computer interaction, and machine learning. The search will focus on candidates applying at the Assistant Professor level, however all qualified applicants will be considered.

The Department of Computer Science has 31 full-time tenured and tenure-track faculty members, 8 research and 6 teaching faculty members, 225 PhD students, over 200 MSE/MSSI students, and over 600 undergraduate students. There are several affiliated research centers and institutes including the Laboratory for Computational Sensing and Robotics (LCSR), the Center for Language and Speech Processing (CLSP), the JHU Information Security Institute (JHU ISI), the Institute for Data Intensive Engineering and Science (IDIES), the Malone Center for Engineering in Healthcare (MCEH), the Institute for Assured Autonomy (IAA), and other labs and research groups. More information about the Department of Computer Science can be found at [www.cs.jhu.edu](http://www.cs.jhu.edu) and about the Whiting School of Engineering at <https://engineering.jhu.edu>.

Applicants should submit a curriculum vitae, a research statement, a teaching statement, three recent publications, and complete contact information for at least three references.

Applications must be made on-line at <http://apply.interfolio.com/78946>. While candidates

who complete their applications by December 15, 2020 will receive full consideration, the department will consider applications submitted after that date. Questions may be directed to [fsearch2020@cs.jhu.edu](mailto:fsearch2020@cs.jhu.edu).

The department is conducting a broad and inclusive search and is committed to identifying candidates who through their research, teaching and service will contribute to the diversity and excellence of the academic community. More information on diversity and inclusion in the department is available at <https://www.cs.jhu.edu/diversity/>.

The Johns Hopkins University is committed to equal opportunity for its faculty, staff, and students. To that end, the university does not discriminate on the basis of sex, gender, marital status, pregnancy, race, color, ethnicity, national origin, age, disability, religion, sexual orientation, gender identity or expression, veteran status or other legally protected characteristic. The university is committed to providing qualified individuals access to all academic and employment programs, benefits and activities on the basis of demonstrated ability, performance and merit without regard to personal factors that are irrelevant to the program involved.

### **Trinity College Computer Science Department Assistant Professor**

Applications are invited for a tenure-track position in computer science at the rank of Assistant Professor to start in the fall of 2021. Candidates

must hold a Ph.D. in computer science at the time of appointment. We are seeking candidates with teaching and research interests in applied areas associated with data analytics, such as database and information systems, data mining and knowledge discovery, machine learning, and artificial intelligence, but other related areas will also be seriously considered.

Trinity College is a coeducational, independent, nonsectarian liberal arts college located in, and deeply engaged with, Connecticut's capital city of Hartford. Our approximately 2,200 students come from all socioeconomic, racial, religious, and ethnic backgrounds across the United States, and seventeen percent are international. We emphasize excellence in both teaching and research, and our intimate campus provides an ideal setting for interdisciplinary collaboration. Teaching load is four courses per year for the first two years and five courses per year thereafter, with a one-semester leave every four years. We offer a competitive salary and benefits package, plus a start-up expense fund. For information about the Computer Science Department, visit: <http://www.cs.trincoll.edu/>

Applicants should submit a curriculum vitae and teaching and research statements and arrange for three letters of reference to be sent to: <https://trincoll.peopleadmin.com/>

Consideration of applications will begin on December 15, 2020, and continue until the position is filled.

Trinity College is an Equal-Opportunity/Affirmative-Action employer. Women and members of minority groups are encouraged to apply.



### **Department of Electrical and Computer Engineering Graduate School of Engineering and Management Air Force Institute of Technology (AFIT) Dayton, Ohio Faculty Position**

The Department of Electrical and Computer Engineering at the Air Force Institute of Technology is seeking applications for a tenured or tenure-track faculty position. All academic ranks will be considered. Applicants must have an earned doctorate in Electrical Engineering, Computer Engineering, Computer Science, or a closely affiliated discipline by the time of their appointment (anticipated 1 September 2021).

We are particularly interested in applicants specializing in one or more of the following areas: autonomy, artificial intelligence / machine learning, navigation with or without GPS, cyber security, and VLSI. Candidates in other areas of specialization are also encouraged to apply. This position requires teaching at the graduate level as well as establishing and sustaining a strong DoD relevant externally funded research program with a sustainable record of related peer-reviewed publications.

The Air Force Institute of Technology (AFIT) is the premier Department of Defense (DoD) institution for graduate education in science, technology, engineering, and management, and has a Carnegie Classification as a High Research Activity Doctoral University. The Department of Electrical and Computer Engineering offers accredited M.S. and Ph.D. degree programs in Electrical Engineering, Computer Engineering, and Computer Science as well as an MS degree program in Cyber Operations.

Applicants must be U.S. citizens. Full details on the position, the department, applicant qualifications, and application procedures can be found at <http://www.afit.edu/ENG/>. Review of applications will begin on 4 January 2021. The United States Air Force is an equal opportunity, affirmative action employer.



# The Essentials of Modern Software Engineering

*Free the Practices from the Method Prisons!*

This text/reference is an in-depth introduction to the systematic, universal software engineering kernel known as “Essence.” This kernel was envisioned and originally created by Ivar Jacobson and his colleagues, developed by Software Engineering Method and Theory (SEMAT) and approved by The Object Management Group (OMG) as a standard in 2014. Essence is a practice-independent framework for thinking and reasoning about the practices we have and the practices we need. **It establishes a shared and standard understanding of what is at the heart of software development. Essence is agnostic to any particular methods, lifecycle independent, programming language independent, concise, scalable, extensible, and formally specified.** Essence frees the practices from their method prisons.

## HIGH PRAISE FOR THE ESSENTIALS OF MODERN SOFTWARE ENGINEERING

“Essence is an important breakthrough in understanding the meaning of software engineering. It is a key contribution to the development of our discipline and I’m confident that this book will demonstrate the value of Essence to a wider audience. It too is an idea whose time has come.” – Ian Somerville, St. Andrews University, Scotland (author of *Software Engineering, 10th Edition*, Pearson)

“What you hold in your hands (or on your computer or tablet if you are so inclined) represents the deep thinking and broad experience of the authors, information you’ll find approachable, understandable, and, most importantly, actionable.”  
– Grady Booch, IBM Fellow, ACM Fellow, IEEE Fellow, BCS Ada Lovelace Award, and IEEE Computer Pioneer

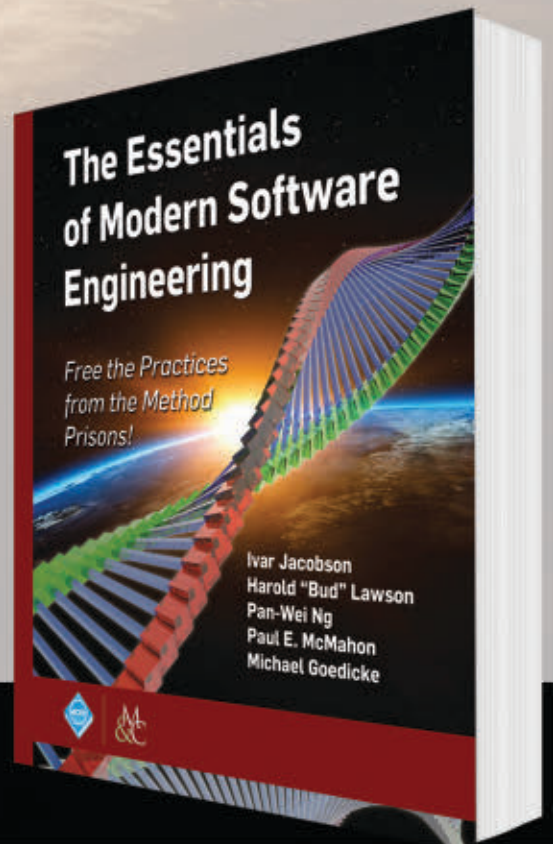
**Ivar Jacobson, Harold “Bud” Lawson,  
Pan-Wei Ng, Paul E. McMahon,  
Michael Goedicke**

ISBN: 978-1-947487-24-6

DOI: 10.1145/3277669

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



**ACM BOOKS**  
Collection I

[CONTINUED FROM P. 160] program-  
ming, even for very young children.

One of the remaining challenges is that we still do not have a solid pipeline of teachers who can meet the rising demand for CS in public schools. We have also not succeeded in making access to CS learning truly equitable. We've worked very hard as a community to focus on diversity, and we're making some gains in terms of gender, race, and ethnicity, but they are not sufficient. I also think that poverty is a bigger and more complex issue. When we talk about equity, very rarely do we talk about socioeconomic issues. We don't like to look at poverty—and I don't just mean in the CS education community, I mean broadly.

**What, in your opinion, have been some of the more effective strategies for encouraging diversity?**

There are two different schools of thought, and both are valid. One school of thought is to change the curriculum to make it more accessible and engaging to everyone—this is impacting undergraduate education as well as elementary and secondary schools. The other focus has been to change the culture in which the learning takes place. Things like what kind of visuals are in the lab, the language we use to address our students, and whether our classroom culture is more competitive or more collaborative have an impact on whether and which students believe they belong.

**What about the challenge of fostering a pipeline of qualified CS teachers?**

One of the realities of teacher education is that it is standards-driven. The job of teacher education programs is to prepare teachers to address and achieve state-level learning standards. The other driving factor is certification. So, without standards and without a pathway to certification, there is absolutely no incentive for teacher preparation programs to prepare CS teachers. I think we've seen amazing progress on the standards side, and some states are now starting to address the certification pathway issue because they see that, finally, there is significant demand. But we are still only seeing small pockets of innovation.

**“Assessment is one of the areas in which I experience a lot of cognitive and emotional dissonance.”**

**Let's talk about your work with Google, which has launched CS programs like CS4HS, one of the earliest efforts to support the professional development of computer science teachers.**

CS4HS preceded my time at Google. In 2008, it was a truly innovative and necessary program. Today, however, there are many groups who are providing professional development for teachers, from Code.org to Mobile CSP. CS4HS is no longer as necessary, so we've transitioned our focus to supporting rigorous CS education research. CS education does not have the deep and wide body of knowledge that other disciplines can rely on, so two years ago, Google launched a program called Computer Science Education Research grants, or CS-ER, through which we provide one-year grants to support innovative research directed at improving teaching and learning in CS in K-12.

**What kinds of proposals have come in thus far?**

The proposals have been hugely diverse and very rigorous, which is great. We've funded projects that looked at the needs of students in rural areas and how to address them, projects that relate to preparing teachers for new certification exams, and projects that focus on the development of curricular material and computational thinking to be introduced to teachers in their pre-service education programs.

**Let's talk about the issue of assessment, which you've commented on before.**

Assessment is one of the areas in which I experience a lot of cognitive and emotional dissonance. At

Google, we frequently get asked to develop assessments to measure student learning, but I believe that teachers always provide the best assessments. I particularly struggle with high-stakes testing, because while I understand that it originally came from a place of trying to ensure equitable learning for all students, it has become all stick and no carrot. I also see how it affects teachers who feel compelled to teach to the test and students who feel tested to death. Unfortunately, I do not find myself capable of articulating a solution.

**Are there other challenges, new or old, that you feel don't get as much attention as they deserve?**

When I was with CSTA, we started with the easier problems and worked our way up to the hardest ones. The easier problems were things like creating resources for teachers, addressing teacher isolation, building a community of teachers, and providing ways for those teachers to grow as leaders. Then we moved on to standards, which was harder, but we achieved it. We have also made enormous strides in helping the public understand why CS education is relevant and necessary.

**Now it's onto the much more complicated issues of access and equity, and ensuring we have a continuing pipeline of CS teachers.**

The challenges that remain are significant, and they are going to take a lot of hard work, but I feel that we're in a space now where all the people who need to be engaged are engaged, including parents, and I'm very, very hopeful. The one little warning bell that rings in my head is that I feel we have about three years to prove that we were right—that students can learn this, that they can learn it effectively, and that it will help them in their futures. Now that we're in the implementation phase, we have to be even more attentive and rigorous in our thinking and our actions to ensure that we're doing the best thing for all students.

*Leah Hoffmann* is a technology writer based in Piermont, NY, USA.

© 2020 ACM 0001-0782/20/11 \$15.00



## Q&amp;A

# Tackling the Challenges of CS Education

*Chris Stephenson on the complex challenges that continue to plague the computer science education community.*

CHRIS STEPHENSON ISN'T afraid to tackle complex problems. The founding Executive Director of the Computer Science Teachers Association (CSTA), current head of Computer Science (CS) Education Strategy at Google, and recipient of the 2018 Outstanding Contributor to ACM Award, Stephenson has worked tirelessly since the late 1980s to advance computer science education at the K-12 level. Here, she talks to us about the challenges that the CS education community still faces, from building a pipeline of qualified CS teachers to ensuring equitable access to learning for all students.

**You've been involved with CS education for decades, but your career path was somewhat winding.**

I've probably already had several careers. I started out working as a radio news broadcaster, and then I moved into public television, where I was a researcher for a public affairs show. In that era, personal computers were just coming into use, and I was fortunate enough to be given a computer at work. It opened my eyes to a world of possibilities. After that, I began working as a technical writer, and eventually I was hired by the compiler writing team at the University of Toronto. That's when my true career in computer science education began. At the time, the university was promoting the use of a programming language called Turing, which they'd developed and were using to teach



introductory courses. They wanted to make it broadly available to high school teachers and students. My efforts to understand how to achieve this gave me a much better understanding of the complexities of formal education.

**In 2004, shortly after you began working toward your Ph.D. in education at Oregon State University, ACM hired you on a part-time basis to start the Computer Science Teachers Association.**

Starting a new organization was an exciting opportunity. Yet getting anyone interested in computer science, at that time, was a tremendous challenge.

**What were some of the specific issues you faced?**

There were a couple of factors impacting the situation for CS educa-

**Working to launch CSTA was exciting, but "Getting anyone interested in computer science ... was a tremendous challenge."**

tion. First, there wasn't any kind of broad public understanding of CS education and its potential place in the canon. Also, many of the original CS school programs were disintegrating with the retirement of a generation of teachers, and CS certainly wasn't on the radar of politicians.

**A lot has changed since then—which is not to say there aren't still challenges.**

It's been really exciting to see the sea change that has happened. Fifteen years ago, if you asked anyone about what programming language they were using and why, they would talk about industrial relevance. When you're dealing with kids who are at least eight years away from employment in the field, that's not an academically sound rationale. Now, there are so many accessible tools to teach [CONTINUED ON P. 159]

# Providing Sound Foundations for Cryptography

*On the work of Shafi Goldwasser and Silvio Micali*

Cryptography is concerned with the construction of schemes that withstand any abuse. A cryptographic scheme is constructed so as to maintain a desired functionality, even under malicious attempts aimed at making it deviate from its prescribed behavior. The design of cryptographic systems must be based on firm foundations, whereas ad hoc approaches and heuristics are a very dangerous way to go. These foundations were developed mostly in the 1980s, in works that are all co-authored by Shafi Goldwasser and/or Silvio Micali. These works have transformed cryptography from an engineering discipline, lacking sound theoretical foundations, into a scientific field possessing a well-founded theory, which influences practice as well as contributes to other areas of theoretical computer science.

This book celebrates these works, which were the basis for bestowing the 2012 A.M. Turing Award upon Shafi Goldwasser and Silvio Micali. A significant portion of this book reproduces some of these works, and another portion consists of scientific perspectives by some of their former students. The highlight of the book is provided by a few chapters that allow the readers to meet Shafi and Silvio in person. These include interviews with them, their biographies and their Turing Award lectures.

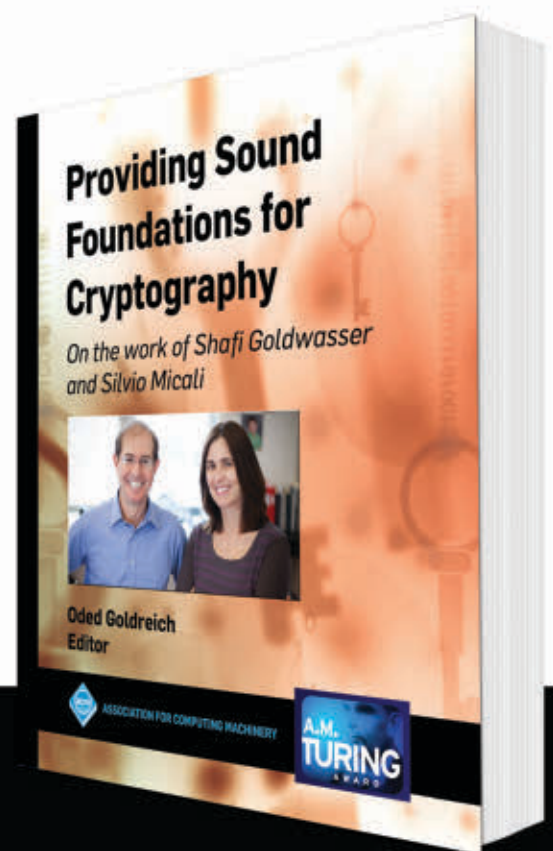
**Oded Goldreich, Editor**

ISBN: 978-1-4503-7267-1

DOI: 10.1145/3335741

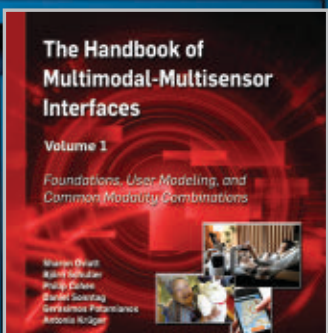
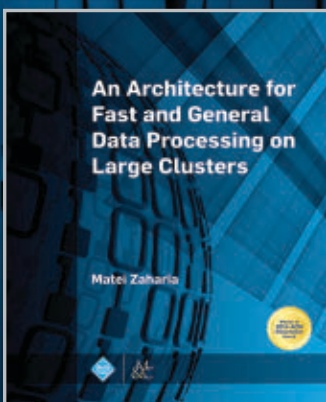
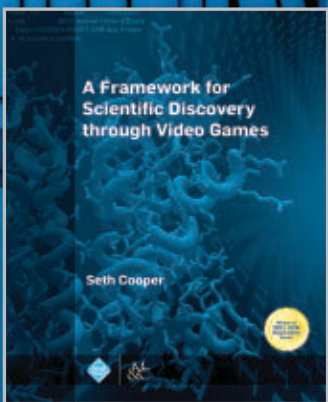
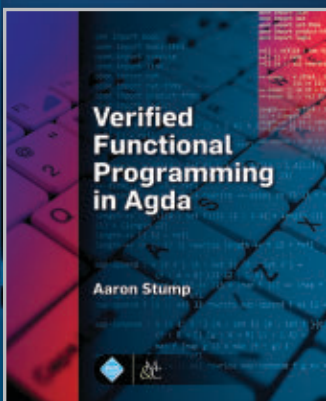
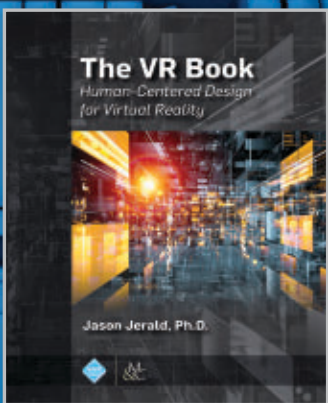
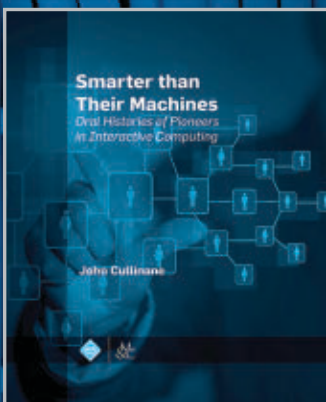
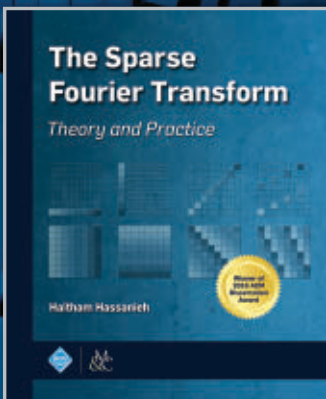
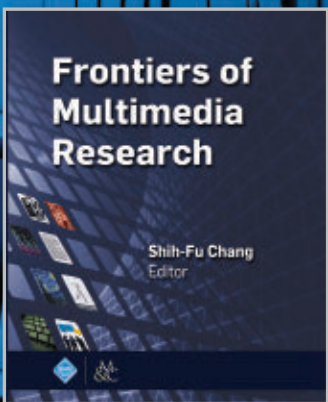
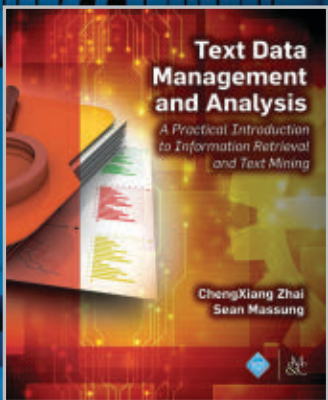
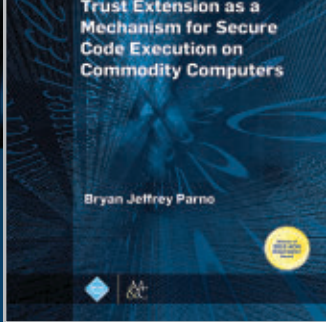
<http://books.acm.org>

<http://store.morganclaypool.com/acm>



**ACM BOOKS**  
Collection II





# 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](mailto:acmbooks-info@acm.org)