

# COMMUNICATIONS

OF THE

# ACM

CACM.ACM.ORG

02/2021 VOL.64 NO.02



## AZERTY amélioré: Computational Design on a National Scale

Let's Not Dumb Down  
the History of  
Computer Science

Semantic Web:  
A Review of the Field

Driving the Cloud to  
True Zero Carbon

Keeping Science on Keel  
When Software Moves

Association for  
Computing Machinery

acm





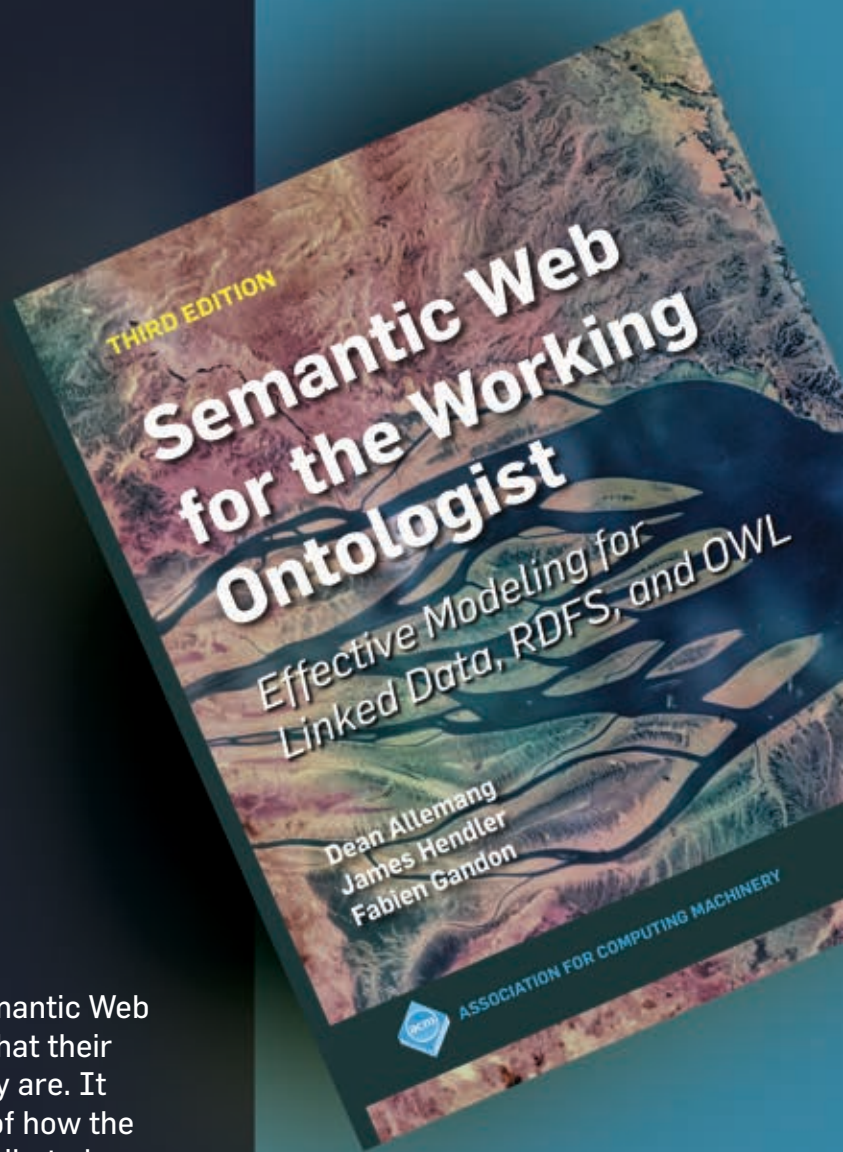
**ACM BOOKS**  
Collection II

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



Athens, Worldwide  
**26-30th June**

# Global meets local

Designing with and for children in a changing world

## ACM SIGCHI Interaction Design & Children Conference

IDC is the premier international conference for researchers, educators and practitioners to share the latest research findings, innovative methodologies and new technologies in the areas of inclusive child-centered design, learning and interaction.

IDC 2021 will be held on **June 26-30, in a hybrid format**, pandemic permitting (online and in Athens, Greece).

This year's theme is: "Global meets local: designing with and for children in a changing world".

## IDC 2021 welcomes the following submissions:

### Full and Short Papers

Abstract: January 25 | Full submission: February 1

### Workshop Organizer Submissions

Abstract: February 1 | Proposal: February 8

### Course Organizer Submissions

March 1

### Children's Design Challenge

March 1 (draft ideas) & May 24 (video submission)

### Doctoral Consortium

March 29

### Work in Progress

April 12



[idc.acm.org/2021](https://idc.acm.org/2021)

We look forward to seeing you at IDC 2021!

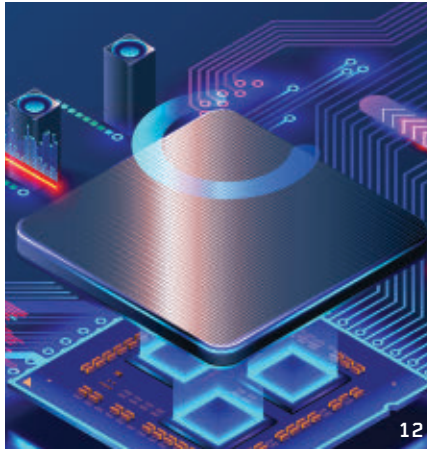
## Departments

- 5 **Editor's Letter**  
**Driving the Cloud to True Zero Carbon**  
*By Andrew A. Chien*
- 
- 7 **Cerf's Up**  
**Half-Baked High-Resolution Referencing**  
*By Vinton G. Cerf*
- 
- 8 **Letters to the Editor**  
**Salary Disputes**
- 
- 10 **BLOG@CACM**  
**Issues Arise When Time Goes Digital**  
Robin K. Hill considers why time can be "a pesky problem for computing."
- 
- 116 **Careers**

## Last Byte

- 120 **Q&A**  
**Bringing Stability to Wireless Connections**  
2020 Marconi Prize recipient Andrea Goldsmith on MIMO technologies, millimeter-wave communications, and her goals as the new dean of Princeton University's School of Engineering and Applied Science.  
*By Leah Hoffmann*

## News



- 12 **Moore's Law: What Comes Next?**  
Moore's Law challenges point to changes in software.  
*By Chris Edwards*
- 
- 15 **The State of Virtual Reality Hardware**  
Advances in VR hardware could finally take the technology mainstream.  
*By Logan Kugler*
- 
- 17 **Technological Responses to COVID-19**  
Companies are finding new ways to enforce social distancing, clean public spaces, and provide substitutes for human workers.  
*By Keith Kirkpatrick*

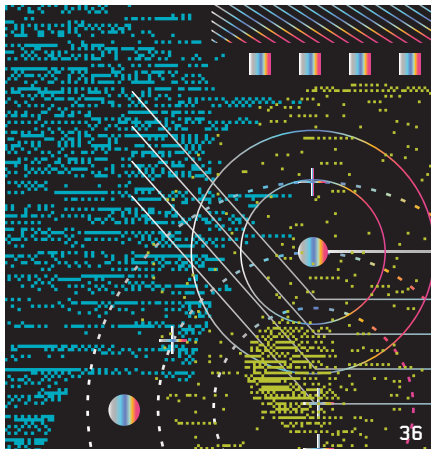
## Viewpoints

- 20 **Economic and Business Dimensions**  
**When Permissioned Blockchains Deliver More Decentralization Than Permissionless**  
Considerations for the governance of distributed systems.  
*By Yannis Bakos, Hanna Halaburda, and Christoph Mueller-Bloch*
- 
- 23 **Education**  
**CAPE: A Framework for Assessing Equity throughout the Computer Science Education Ecosystem**  
Examining both the leading indicators of equity in CS and the lagging indicators of student outcomes.  
*By Carol L. Fletcher and Jayce R. Warner*
- 
- 26 **Kode Vicious**  
**Kabin Fever**  
KV's guidelines for KFH (koding from home).  
*By George V. Neville-Neil*
- 
- 28 **Viewpoint**  
**Cybersecurity: Is It Worse than We Think?**  
Evaluating actual implementations and practices versus stated goals  
*By Chris Maurer, Kevin Kim, Dan Kim, and Leon A. Kappelman*
- 
- 31 **Viewpoint**  
**Polanyi's Revenge and AI's New Romance with Tacit Knowledge**  
Artificial intelligence systems need the wisdom to know when to take advice from us and when to learn from data.  
*By Subbarao Kambhampati*
- 
- 33 **Viewpoint**  
**Let's Not Dumb Down the History of Computer Science**  
Donald Knuth on the best way to recognize the history of computer science.  
*By Donald E. Knuth*





## Practice



- 36 **Differential Privacy: The Pursuit of Protections by Default**  
A discussion with Miguel Guevara, Damien Desfontaines, Jim Waldo, and Terry Coatta

- 44 **The Time I Stole \$10,000 from Bell Labs**  
Why DevOps encourages us to celebrate outages.  
By Thomas A. Limoncelli



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



**About the Cover:**  
A new French keyboard standard AZERTY amélioré is the first designed with the help of computational methods. This month's cover story explores its creation and the methods used to support stakeholder participation in large-scale design projects. Cover illustration by Matt Herring.

## Contributed Articles



- 48 **AZERTY amélioré: Computational Design on a National Scale**  
A new French keyboard standard is the first designed with the help of computational methods.  
By Anna Maria Feit, Mathieu Nancel, Maximilian John, Andreas Karrenbauer, Daryl Weir, and Antti Oulasvirta



Watch the authors discuss this work in the exclusive *Communications* video.  
<https://cacm.acm.org/videos/azerty-ameliore>

- 59 **GDPR Anti-Patterns**  
How design and operation of modern cloud-scale systems conflict with GDPR.  
By Supreeth Shastri, Melissa Wasserman, and Vijay Chidambaram

- 66 **Keeping Science on Keel When Software Moves**  
An approach to reproducibility problems related to porting software across machines and compilers.  
By Dong H. Ahn, Allison H. Baker, Michael Bentley, Ian Briggs, Ganesh Gopalakrishnan, Dorit M. Hammerling, Ignacio Laguna, Gregory L. Lee, Daniel J. Milroy, and Mariana Vertenstein

## Review Articles

- 76 **A Review of the Semantic Web Field**  
Tracing the triumphs and challenges of two decades of Semantic Web research and applications.  
By Pascal Hitzler



Watch the author discuss this work in the exclusive *Communications* video.  
<https://cacm.acm.org/videos/semantic-web>

- 84 **DP-Cryptography: Marrying Differential Privacy and Cryptography in Emerging Applications**  
Synthesizing the emerging directions of research at the intersection of differential privacy and cryptography.  
By Sameer Wagh, Xi He, Ashwin Machanavajhala, and Prateek Mittal

## Research Highlights

- 96 **Technical Perspective**  
**Programming Microfluidics to Execute Biological Protocols**  
By Nada Amin
- 97 **BioScript: Programming Safe Chemistry on Laboratories-on-a-Chip**  
By Jason Ott, Tyson Loveless, Chris Curtis, Mohsen Lesani, and Philip Brisk
- 105 **Technical Perspective**  
**Solving the Signal Reconstruction Problem at Scale**  
By Zachary G. Ives

- 106 **Scalable Signal Reconstruction for a Broad Range of Applications**  
By Abolfazl Asudeh, Jeess Augustine, Saravanan Thirumuruganathan, Azade Nazi, Nan Zhang, Gautam Das, and Divesh Srivastava



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**

Iliia Rodriguez

**Columnists**

David Anderson; Michael Cusumano;

Peter J. Denning; Mark Guzdial;

Thomas Haigh; Leah Hoffmann; Mari Sako;

Pamela Samuelson; Marshall Van Alstyne

**CONTACT POINTS**

**Copyright permission**

permissions@hq.acm.org

**Calendar items**

calendar@cacm.acm.org

**Change of address**

acmhelp@acm.org

**Letters to the Editor**

letters@cacm.acm.org

**WEBSITE**

http://cacm.acm.org

**WEB BOARD**

**Chair**

James Landay

**Board Members**

Marti Hearst; Jason I. Hong;

Jeff Johnson; Wendy E. MacKay

**AUTHOR GUIDELINES**

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

**ACM ADVERTISING DEPARTMENT**

1601 Broadway, 10<sup>th</sup> Floor

New York, NY 10019-7434 USA

T (212) 626-0686

F (212) 869-0481

**Advertising Sales Account Manager**

Iliia 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**

Virgilio Almeida; Terry Benzel; Michael L. Best;

Judith Bishop; Lorrie Cranor; Boi Falting;

James Grimmelmann; Mark Guzdial;

Haym B. Hirsch; Anupam Joshi; Richard Ladner;

Carl Landwehr; Beng Chin Ooi; Francesca Rossi;

Len Shustek; Loren Terveen; Marshall Van

Alstyne; Jeannette Wing; Susan J. Winter

**PRACTICE**

**Co-Chairs**

Stephen Bourne and Theo Schlossnagle

**Board Members**

Eric Allman; Samy Bahra; Peter Bailis;

Betsy Beyer; Terry Coatta; Stuart Feldman;

Nicole Forsgren; Camille Fournier;

Jessie Frazelle; Benjamin Fried; Tom Killalea;

Tom Limoncelli; Kate Matsudaira;

Marshall Kirk McKusick; Erik Meijer;

George Neville-Neil; Jim Waldo;

Meredith Whittaker

**CONTRIBUTED ARTICLES**

**Co-Chairs**

James Larus and Gail Murphy

**Board Members**

Robert Austin; Kim Bruce; Alan Bundy;

Peter Buneman; Premkumar T. Devanbu;

Jane Cleland-Huang; Yannis Ioannidis;

Trent Jaeger; Somesh Jha; Gal A. Kaminka;

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

Hannes Werthner; Reinhard Wilhelm;

Rich Wolski

**RESEARCH HIGHLIGHTS**

**Co-Chairs**

Shriram Krishnamurthi

and Orna Kupferman

**Board Members**

Martin Abadi; Amr El Abbadi;

Animashree Anandkumar; Sanjeev Arora;

Michael Backes; Maria-Florina Balcan;

Azer Bestavros; David Brooks; Stuart K. Card;

Jon Crowcroft; Lieven Eeckhout;

Alexei Efron; Bryan Ford; Alon Halevy;

Renot Heiser; Takeo Igarashi;

Srinivasan Keshav; Sven Koenig;

Ran Libeskind-Hadas; Karen Liu;

Tim Roughgarden; Guy Steele, Jr.;

Robert Williamson; Margaret H. Wright;

Nicholai Zeldovich; Andreas Zeller

**SPECIAL SECTIONS**

**Co-Chairs**

Sriram Rajamani, Haibo Chen,

and P. J. Narayanan

**Board Members**

Sue Moon; Tao Xie; Kenjiro Taura; David Padua

**ACM Copyright Notice**

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

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

**Subscriptions**

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

**ACM Media Advertising Policy**

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

**Single Copies**

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

**COMMUNICATIONS OF THE ACM**

(ISSN 0001-0782) is published monthly by ACM Media, 1601 Broadway, 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







Andrew A. Chien

DOI:10.1145/3445037

# Driving the Cloud to True Zero Carbon

**T**HE RIGHT VISION is to operate the cloud with zero-carbon emission from power (scope 2). Not just offsetting through renewable energy purchases. Not just 24x7 matching. True zero carbon in electric power consumed, and with no increase as the cloud continues to grow. That's the right vision for our proud computing technology community to lead the fight against climate change, and to see increasing use of computing as a positive force to slow climate change.<sup>a,b</sup>

**Why must we act?** The power grid is decarbonizing, but progress is slow. Aggressive states (for example, California and New York) have zero-carbon goals 20 or more years in the future, 2045 and 2040. Nationally, the U.S. produced 19% of its electric power from renewable resources (2020), and with "datacenter alley" reporting 12% renewables<sup>c</sup> (Northern Virginia). This trails the world's 26% renewables today, and U.S. renewables are projected to double to 38% by 2050. At that rate, full decarbonization may be a century away!<sup>d</sup> Substantial progress toward zero carbon in the next 10 years depends on aggressive action, cloud computing cannot just depend on power grid decarbonization.

In 2020, the cloud's power consumption exceeded 2% of total U.S. power, and hyperscale providers exceed 6% in regional power markets and grids.<sup>e</sup> Add to that annual growth of 28% (2017–2019)<sup>f</sup> as well as acceleration from COVID-spurred digitalization and machine learning, and it's clear that cloud power consumption

cannot be ignored in serious attempts to reduce anthropogenic carbon emissions.

**Is it possible?** Leading computing companies—notably Google, Facebook, and Apple have been carbon neutral for years. Amazon has committed to becoming carbon neutral by 2030. An avalanche of more aggressive commitments has been made in 2020.<sup>g</sup> Why are these computing giants moving now?

Climate-change natural disasters have become common. The disruption of floods and wildfires has led hedge funds to shift climate-risk from a second to "first ledger" issue, and they have moved to assess climate-risk in business valuation.<sup>h</sup> Climate change has come to the fore as a business concern, and this compels the business leaders of computing technology companies to make stronger, more ambitious carbon-reduction commitments, and to be on the side of progress toward true zero carbon.

In 2020, we have seen major national commitments to carbon-neutral economies by Japan (2050) and China (2060),<sup>i</sup> joining the European Union (2045). Government commitments produce growing pressure on all of the economy. And, with regulation of "big tech" on deck, that contribution to economic growth no longer confers a free pass.<sup>j</sup>

So, yes, it's possible. The public, governments, and the hedge funds are all aligned. These commitments acknowledge responsibility and create a growing economic drive. There's little doubt we have the technological capability. The technical challenges are around how to

do it as cheaply as possible. Solving these challenges requires new research, technology, and large-scale investment.

**What must we do?** Here's a roadmap.

- ▶ Learn about renewables and the modern power grid: the key to carbon emissions is where and when power is consumed—not just power efficiency

- ▶ Create applications that can flex when and where they consume power, enabling time and space shifting

- ▶ Create carbon-aware applications that exploit flexibility and carbon-content information to reduce carbon emissions

- ▶ Design novel hardware architectures (and datacenter facilities) that provide inexpensive capacity to support such workload shifting

We must redesign cloud software and hardware to flexibly follow renewable energy. For cloud computing, the majority of carbon emissions arise from power consumed during operation (80% for typical four-year use). But embodied carbon for hardware and datacenter infrastructure (scope 3) cannot be ignored.<sup>k</sup> One effective way to do this is to extend the lifetime of computing hardware, and creating a circular ecosystem.<sup>l</sup>

Let's all drive cloud computing to true zero carbon!

<sup>j</sup> A. Satariano. Big fines and strict rules unveiled against 'big tech' in Europe, *NYTimes* (Dec. 15, 2020).

<sup>k</sup> B. Manne. Architecting a Sustainable Planet. Keynote at IEEE MICRO-53 Conf. Oct. 2020.

<sup>l</sup> Extending the lifetime of scientific computing equipment; <http://bit.ly/3mBG0XG/>, and ITRenew: Expect more from your IT hardware; <https://www.itrenew.com/>

**Andrew A. Chien**, EDITOR-IN-CHIEF  
*COMMUNICATIONS OF THE ACM*

**Andrew A. Chien** is the William Eckhardt Distinguished Service Professor in the Department of Computer Science at the University of Chicago, Director of the CERES Center for Unstoppable Computing, and a Senior Scientist at Argonne National Laboratory.

Copyright held by author/owner.

<sup>a</sup> A.A. Chien. Owning computing's environmental impact. *Commun. ACM*, Mar. 2019.

<sup>b</sup> A.A. Chien. What do DDT and computing have in common? *Commun. ACM*, June 2020.

<sup>c</sup> Dominion Energy; <http://bit.ly/3h8OZd1>

<sup>d</sup> Energy Information Agency. Annual Energy Outlook 2020, (Jan. 29, 2020); <https://www.eia.gov>.

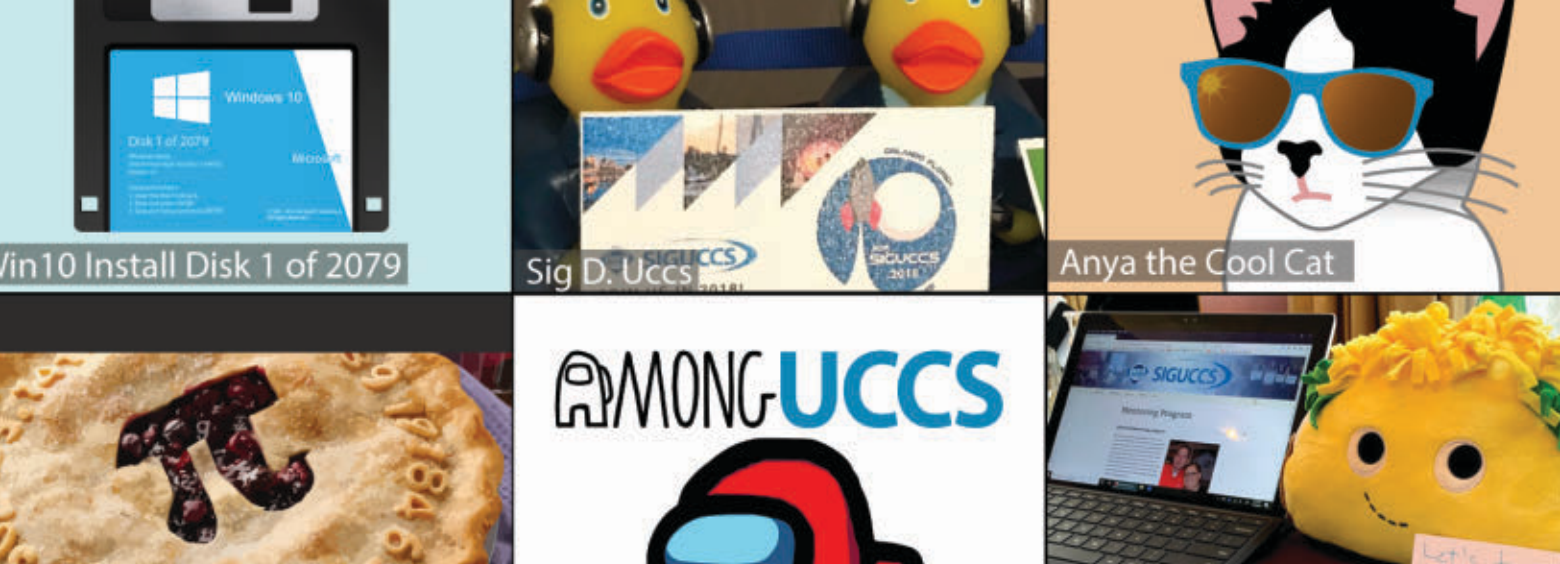
<sup>e</sup> Dominion Energy; <https://bit.ly/2Kw1X8L>

<sup>f</sup> Google. Realizing a carbon-free future: Google's third decade of climate action, Sept. 2020.

<sup>g</sup> S. Pichai (Google) 24x7 matching by 2030, J. Bezos (Amazon) Carbon Neutral by 2040, and S. Nadella (Microsoft) Lifetime offset by 2050.

<sup>h</sup> M. Peregrine. Blackrock heats up climate change pressure on boards. *Forbes*, (July 19, 2020).

<sup>i</sup> China's next economic transformation: Going carbon neutral by 2060. *WSJ*, (Oct. 29, 2020). Japan promises to be carbon neutral by 2050. *Economist*, (Oct. 29, 2020).



# SIGUCCS 48th Annual Conference

March 14 - April 30, 2021 | Online

Register now and urge your colleagues in higher education IT support to do the same! This year's conference will be **FREE** for attendees and conducted virtually. Join us online for sessions between March 14 and April 30th. Register for sessions featuring your colleagues' work over the past year, social events and our plenary speakers, Margaret Burnett and Susan Zvacek.



**Registration is Open!**

We're looking forward to seeing what's possible in our virtual seminar series including sessions on:

- Lessons Learned from Rapid Changes in Teaching Modality
- Humanity in the Workplace
- Technology
- Leadership
- And MORE...

Plus pre-conference seminars and the newly virtualized poster session along with our ever popular lightning talks!



**Register for free:**

<http://tiny.cc/SIGUCCSRegister>

ACM SIGUCCS is the Special Interest Group on University and College Computing Services





Vinton G. Cerf

DOI:10.1145/3442186

# Half-Baked High-Resolution Referencing

**I**N THE PAST, I have written about digital preservation. I would like to turn to a related topic that I will call *high-resolution referencing*. In conventional print publication media, it is possible to cite books, chapters, papers, sections, pages, paragraphs, and even sentences. One reason this is possible is that these media *fix* the work indelibly. Of course, one must have the correct version of the publication in hand, so to speak, since pagination is a function of font size, for example. In the World Wide Web, the Hypertext Transport Protocol and the Hypertext Markup Language serve the needs of users to refer to Web pages and can do so with considerable precision by using features of extended URLs to reference specific sections of Web pages. URLs referencing anchor points within a Web page offer what I will refer to as a high-resolution reference. Of course, if the Web page has been changed, such references may fail with the too familiar “404 page not found” or similar error message.

In the world of Google Docs, and other document processing systems, it is often possible to keep track of the time sequence in which edits have been made so as to “undo” an action or to return to a previous version of the document. This leads me to wonder whether time resolution, in addition to space resolution, might be an interesting functionality to instantiate. A reason this may be of interest is Web page references are beginning to show up in print and other media with the annotation “retrieved <date>” included. While this information is helpful, a later reader may not find what the reference intended if the Web page has evolved since it


was referenced by the writer. One might imagine a construct in which the document (Web page, PDF?) includes timestamped edit information such that the version of the document at a given date/time might be reconstructed. Since editing can be a messy process, one supposes the writer, interested in capturing versions, might want to identify at what point a document should be “versioned.” This is not unlike existing mechanisms for keeping track of software versions by “checking out” and “checking in” versions of source code. This could become metadata for the document in the same sort of way that breakpoints and periodic backups allow for recovery to a known condition in a lengthy computation.

Assuming for a moment that this would be an interesting capability, it remains to figure out how to imple-

**I wonder whether time resolution, in addition to space resolution, might be an interesting functionality to instantiate.**

ment it for various cases. In the case of Google Docs, the internal representation appears to allow the document to be reconstructed in its entirety upon fetching, from its initial instantiation and subsequent editing. This suggests a versioning record could be as simple as recording a date/time at which the document is at “version X” for some value of X. A reference to “version X” of the document would reconstruct all edits up until the date/time at which version X was “marked.” It seems equally feasible to export a document in a variety of formats including Web page HTML including an indication of which version it represents.

It is not clear to me whether one could incorporate such time-based mechanisms within an HTML or PDF document without incurring either overhead for generating and storing every “version” or reconstructing the entire object every time the object is retrieved as happens with Google Docs. Assuming that time or version-based citations are feasible and useful, there comes the question of how to generate the references. Generating these citations sounds like a non-trivial exercise and tools are emerging to assist authors with the generation of citations and for readers to use them. One set of tools created by Frode Hegland and his collaborators can be found at <https://www.augmentedtext.info>.

I am sure readers of this column will have a lot to teach me about floating half-baked ideas. 

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

Copyright held by author.

# Salary Disputes

**I**N MOSHE VARDI'S September 2020 column, "Where Have All the Domestic Graduate Students Gone?," the short but woefully incomplete answer is that the wage premium for a Ph.D. in CS is simply too small to justify foregoing five years of industry-level salary. But why is that the case?

Part of the answer may be due to government policy discussed back in 1989, when an NSF document addressed the "problem" of Ph.D. salaries being too high, and suggested as a remedy increasing the pool of international students (<https://bit.ly/2IuFZl7>). This would swell the labor market, holding down wage growth. The foreign students would receive nonmonetary compensation in the form of a green card:

"A growing influx of foreign Ph.D.'s into U.S. labor markets will hold down the level of Ph.D. salaries to the extent that foreign students are attracted to U.S. doctoral programs as a way of immigrating to the U.S."

But the domestic students would find that the resulting wage suppression would make Ph.D. study a bad choice:

"... a key issue [for the domestic students] is pay. The relatively modest salary premium for acquiring [a] Ph.D. may be too low to attract a number of able potential graduate students ... A number of them will select alternative career paths ... by choosing to acquire a 'professional' degree in business or law ... For these baccalaureates, the effective premium for acquiring a Ph.D. may actually be negative."

**Perhaps we should double the \$\$\$'s/year and double the number of awards in computer science.**

To be sure, it is not fully clear whether this represented official NSF policy. But in any case, the effects predicted did indeed occur in the subsequent years, and we now see university CS departments struggling to find domestic applicants.

Whether justified or not, the recent restrictions placed on international students expose a dangerous dependency on obtaining students from abroad. Many events beyond U.S. control could result in this pool drying up. American institutions must address this urgent issue.

**Norman Matloff**, Davis, CA, USA

### Author's Response:

A "working draft" does not represent government policy. Having watched the evolution of graduate studies in computing over the past 40 years, I am highly skeptical of the argument that the current situation is the result of a directed government policy. As to economic impact of immigration, an authoritative source is a 2017 National Academies report: <https://bit.ly/34o3rJ3> The report concluded "The long-term impact of immigration on the wages and employment of native-born workers overall is very small."

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

### Editor-in-Chief's Response

Domestic U.S. students have extraordinary opportunities in industry, but we have done little to create incentives for their advanced graduate study. The NSF graduate fellowship program provides less annual support than a typical research assistantship, and the number of awardees has not been increased in more than 10 years. Perhaps we should double the \$\$\$'s/year and double the number of awards in computer science. A common complaint is these awards go disproportionately to "top" universities—as coincidentally do "top" students. A remedy would be to distribute these awards over the top 70



## 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)





departments (approximately half the departments reporting in the Taulbee survey). Any such programs would by no means make Ph.D. study competitive with industry pay, but would show our commitment to encouraging such study (and be a step toward the much higher Ph.D. program compensation offered in countries such as Switzerland).

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

### Read Worthy

While reading John MacCormick's Viewpoint "Using Computer Programs and Search Problems for Teaching Theory of Computation" (Oct. 2020, p. 33), I couldn't help but think that the author was describing using Niklaus Wirth's book, *Algorithms + Data Structures = Programs*, in an introductory CS course, much like the one I took in 1983. I wholeheartedly agree. I still refer to that book every once in a while. It's one of the oldest programming books on my shelf.

**Lee Riemenschneider**,  
Lafayette, IN, USA

### Author's Response:

*Lee Riemenschneider's insight helped me view these ideas from a different angle. Niklaus Wirth gave our community a new perspective on programming languages, algorithms, and data structures—a perspective optimized for teaching and learning, not for doing research. Perhaps the approach described in my Viewpoint can do the same for the theory of computation, offering novice students a treatment optimized for learning rather than academic research.*

**John MacCormick**, Carlisle, PA, USA

### Editor-in-Chief's Response:

*A timely highlight of Niklaus' extraordinary work! The March issue of Communications will include a Viewpoint by Nicklaus Wirth reflecting on 50 years of Pascal.*

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

### Lost in Space

Regarding George Neville-Neil's October 2020 Kode Vicious column "Sanity vs. Invisible Markings" (p. 28), writers should pay more attention to

differences in their terminology for "invisible markings."

A "blank" is a single character with its unique Unicode, ASCII, EBCDIC, or other code.

A "space" is a *complete* row of blanks.

A "tab" is a *partial* row of blanks with a length that may be programmable and vary from one "system" to another.

Obviously the "space bar" has been misnamed for more than one hundred years. When someone presses the *space* bar, the result is one *blank*.

**Richard Rosenbaum**,  
Bloomfield Hills, MI, USA

### Everything Old Is New Again

The biggest dark pattern in the Practice article "Dark Patterns" (Sept. 2020, p. 42) is itself. The new discovery they report is just to repackaging of old wine in new bottles. Technologists, in particular, seem to be immune to learning from history. If the authors read Jill Lepore's *These Truths: A History of the United States*, they would learn that psychology in communications has been rediscovered with every new medium. It was learned by newspapers in 1770, in telegraphy in 1850, in radio in 1920 in TV in 1950 and now in the Internet. I can remember my father reading the "Women's Wear Daily" in the 1950s telling me about the academics that had just discovered advertising and wrote articles telling the department stores to advertise in the summer when traffic was slow, rather than near holidays when traffic was already heavy. The authors end up telling design engineers to set standards for themselves because of a "misalignment between the industry and society" without telling us how to discover the needs (wants?) of society. The idea that some neutral third-party advocacy agency would be a stand-in for society sounds like Plato in the Republic asking for society to be ruled by Philosopher kings because democracy was too fragile to survive. That didn't work out too well.

**Tom Jones**, Seattle, WA, USA

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

© 2021 ACM 0001-0782/21/2 \$15.00

## Coming Next Month in COMMUNICATIONS

**The Decline of Computers as General-Purpose Technology**

**Education Inventions and Female Enrollment**

**Gender Trends in CS Authorship**

**Knowledge Graphs**

**Niklaus Wirth on Pascal at 50**

**Cyber Reconnaissance Techniques**

**A Conversation with Werner Vogels**

**Out-of-This-World Additive Manufacturing**

**What Can the Maker Movement Teach Us?**

**Patient Clinic Relationship in AI-based Advice**

**Understanding Deep Learning Requires Rethinking Generalization**

**3D Localization of Sub-Centimeter-Sized Devices**

Plus the latest news about how quantum computing solves math and physics, fact-checking fake news, and avoiding/fixing bias in image recognition.

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:10.1145/3440990

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

## Issues Arise When Time Goes Digital

*Robin K. Hill considers why time can be "a pesky problem for computing."*



**Robin K. Hill**  
**Deadlines of the Digital Turn**  
November 7, 2020  
<https://bit.ly/3mDabcY>

Grading online, I spot a notification from our learning management system displayed on a student's assignment entry:

Submitted:  
Oct 19 at 8:30am LATE

I glance over to the heading of the assignment and see this text:

Exercise #7 Due:  
Oct 19 at 8:30am

Well, that's vexatious, but seen before and easily accommodated by not counting the work late. Because, surely, it's not late! The online instructor help describes this quirk as a feature, not a bug: "... For example, if you set a due date of September 19 at 4:15pm, any student submission made at or after September 19 at 4:15:01 is marked late." So the full minute of interest is not granted. Is this fair? On a high-stakes assignment, a student would have a legitimate objection: Isn't the deadline at the end of the minute, not the start of the minute?

Griping aside, what would we have this system do instead? We intend for the assignment not to be late before 8:31. The programmer, given that specification, could program the expiration at 8:30:59, but that still leaves the gap between that point (8:30:59:00) and the deadline, a gap packed with milliseconds (or other subdivisions such as "jiffies," I see in Wikipedia<sup>1</sup>). A test of the system time for LATE = (hh:mm > next(8:30)) is far-fetched because there is no function next() that computes 8:31 to be the time that comes after 8:30. The instructor doesn't want to say "late at 8:31" anyway, but rather "due at 8:30." The instructor does not want to cross over into the next time unit in order to establish a deadline.

What exactly *does* the instructor want? "You know what I mean: I want the papers in my possession at 8:30, well, right after 8:30." What the instructor wants is a phenomenon of the physical world—a tidy pile of submissions stacked up the day before, with perhaps one or two students racing to the professor's office in the morning, paper in hand, as the hands of the clock approach 8:30. As the

professor sees the submission thrust toward her, she can look at her watch and declare whether the deadline is met, annotating the paper appropriately. She can then shut the door for the grading session, starting right then at 8:30, whatever version of 8:30 she defines. Shifting this scenario to the digital world is not as straightforward as she had expected.

In *Communications*, George Neville-Neil has pointed out time is a pesky problem for computing, in the establishment of synchronization and syntonization, in the design of clock hardware, in the querying of system time, and in just about every other respect.<sup>2</sup> We set aside these interesting issues, as well as those that reach beyond the technical into the social, such as bizarre stock market trends due to lightning-fast high-frequency trading. The philosophical questions include whether time supervenes on events, whether the present is privileged, what sort of formalism is suitable for temporal reasoning, and many other interesting issues,<sup>3</sup> but this is not about those either. This is the problem of designating a particular point on a line in a way that cuts



the values into “before” and “here.” The time just “before” is a block (of the length of whatever unit is in use), a discrete construct, whereas the time “here” is an interstice of length zero.

Consider a significant and well-known time of day: midnight. Suppose I tell my students an assignment is due on a certain date. They know the date ends at midnight, and reasonably infer any clock time bound to that date is acceptable for submission. What is the very last instant that meets the standard? When exactly is the midnight at the end of the day called, say, November 9<sup>th</sup>? Is it at 1200 hours past noon, or is that time actually November 10<sup>th</sup>? Apparently, nobody knows. We can fix the accuracy at the level of seconds, avoiding Zeno’s paradox and making that very last instant our familiar discrete subdivision of a minute. But which second? Is it 23:59:59 or is it 24:00:00? Oh, dear; it’s the point between. If we make our assignments due at noon, do we call it 11:59:59 or 12:00:00? And, if the latter, do we call it “A.M.” or “P.M.”? Oh, dear—an interstice again, and because A.M. means *ante meridiem* and P.M. means *post meridiem*, neither works for the actual meridies.

Let’s turn to the authority of a nation well-versed in timetables, Britain’s National Physical Lab: “To avoid confusion, it is always better to use the 24-hour clock, so that 12:00 is 12 noon. Therefore 24:00 Sunday or 00:00 Monday are both midnight meaning Sunday to Monday.”<sup>4</sup> This authority at least validates the ambiguity. So there is no such thing as midnight on a certain day; there is only the transition between one day and the next. Although midnight is a high-profile time of day, this is a problem manifest only in a digital context. The scheduling of an event such as a train departure or a pagan ceremony is performed on a human scale; someone declares it. The simple change from one date to the next is not necessarily declared, but exposed only by the human need for noting some occurrence before or after the placement of midnight.

That gives us a clue about the root of the problem. It is not the artificial construction of our system of time,

but the digital turn. The deadline of the past allowed people to take care of it in whatever way seemed appropriate, unhampered by any mandate to locate the exact end of a block of time. Even sharp deadlines were enforced by simple human fiat, and still are, in most daily business. Some force other than time itself does the reckoning, and on a continuum that embraces loose placement along the milliseconds. It is the physical manifestation of the deadline that counts, not the deadline itself. The Stock Exchange opens on a bell, and the New Year arrives when observers in Times Square see that the ball has fallen.

This is not novel, but the same problem as designating a point on the continuum using a real number with a finite decimal expansion, a problem that used to be housed in the applications of mathematics. Now that time is discrete, we are trying to force a discrete representation into a continuous phenomenon. This occurs in many other realms as well, of course—distance, volume, anything measurable.<sup>5</sup> We can avoid using 12 A.M. and 12 P.M., but what about those pesky due dates? When I give a deadline, my students trust that I am passing a designation of a block, with the deadline at the end, but that’s not a well-defined type. I will avoid using the end of a day, also known as midnight, as a deadline on a digitally timed platform, and I will try telling my students to submit “before 8:30.”

#### References

1. Wikipedia contributors, “Unit of time.” Wikipedia, The Free Encyclopedia, [https://en.wikipedia.org/wiki/Unit\\_of\\_time](https://en.wikipedia.org/wiki/Unit_of_time), Accessed 7 Nov. 2020.
2. Neville-Neal, G. Time is an Illusion. *Commun. ACM* 59\_1 (Jan. 2016).
3. Dowden, B. No date given. Time. *The Internet Encyclopedia of Philosophy*. ISSN 2161-0002.
4. National Physical Laboratory, Questions and Answers: Is midnight 12am or 12pm? <https://www.npl.co.uk/resources/q-a/is-midnight-12am-or-12pm>
5. Tal, E., “Measurement in Science,” *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (Ed.).

**Robin K. Hill** is a lecturer in the Department of Computer Science and an affiliate of both the Department of Philosophy and Religious Studies and the Wyoming Institute for Humanities Research at the University of Wyoming. She has been a member of ACM since 1978.

## INTERACTIONS



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

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



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

Follow us on Facebook and Twitter



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

Association for Computing Machinery



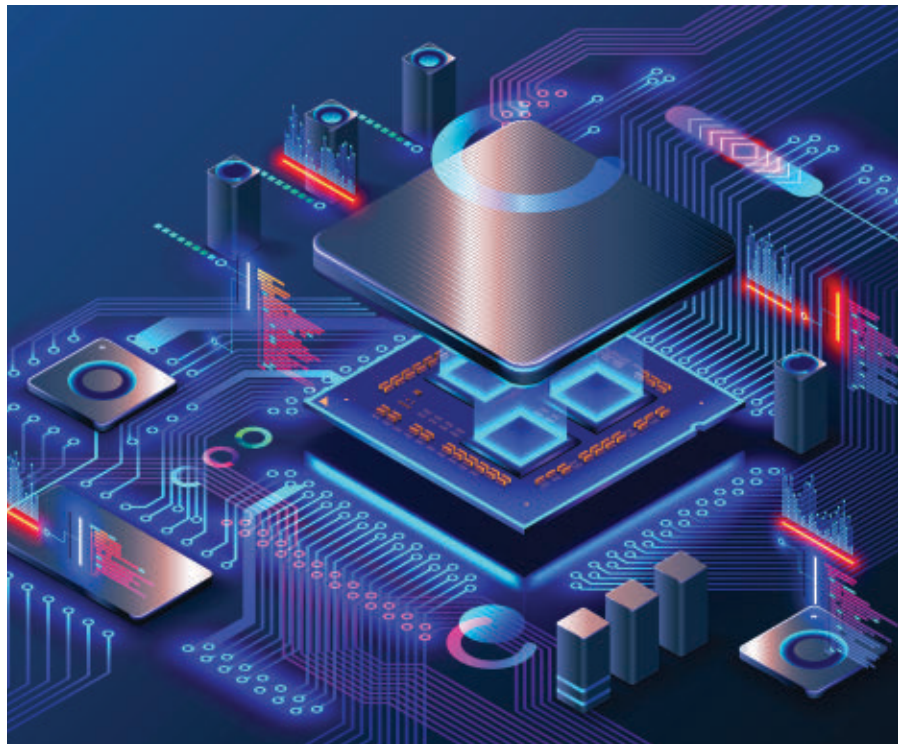
## Moore's Law: What Comes Next?

*Moore's Law challenges point to changes in software.*

**C**OMPUTER DESIGNERS ARE becoming increasingly concerned about the ending of Moore's Law, and what it means for users if the industry can no longer count on the idea that the density of logic circuits will double every two years, as it has for close to half a century. It may mean radical changes to the way users think about software.

Leading researchers in semiconductor design point out that, although logic density is butting up against physical limits, it does not necessarily spell the end of Moore's Law itself. Gordon Moore's speech at the 1975 International Electron Device Meeting (IEDM) predicted significant increases in chip size and improvements in circuit design as part of the scaling process, in addition to regular reductions in transistor size and interconnect spacing.

During a September virtual meeting of the IEEE International Roadmap for Devices and Systems group, chairman and Intel director of technology strategy Paolo Gargini, argued, "Though Gordon made this clear, people have concentrated only on dimensional scaling. That's the reason why people have doubts about the next technology nodes. It appears as though we are in a



crisis, but we are not, because of the other two components."

"Circuit cleverness" as described by Moore in his 1975 speech, has made a strong contribution in recent years. Philip Wong, professor of electrical engineering at Stanford University, says greater cooperation between circuit

designers and the engineers who work on the core process technology has made it possible to eke more gains out of each new node than would be possible using just dimensional scaling. Advances such as burying power rails under transistors and stacking transistors should continue to provide some gains

for perhaps two or three generations, out to the latter half of this decade. The remaining directions for future improvements at the physical level are to build out in terms of area by adding more layers of logic gates or other devices. Some warn, however, that this direction has its own imitations.

Neil Thompson, a research scientist at the Massachusetts Institute of Technology (MIT), says, “When you look at 3D (three-dimensional) integration, there are some near-term gains that are available. But heat-dissipation problems get worse when you place things on top of each other.

“It seems much more likely that this will turn out to be similar to what happened with processor cores. When multicore processors appeared, the promise was to keep doubling the number of cores. Initially we got an increase, and then got diminishing returns.”

One option is to make more efficient use of the available transistor count. In the lecture to commemorate their 2017 ACM A.M. Turing Award, John Hennessy and David Patterson argued there is a rich vein to mine in highly specialized accelerators that dispense with the heavy overhead of general-purpose computing, much of it due to highly wasteful memory accesses caused by repeated instruction and data fetches, as a way of providing the performance that Moore’s Law may not be able to support.

Paul Kelly, professor of software technology at Imperial College, London, uses the term “Turing tariff” to refer to the cost of performing functions using general-purpose hardware. The term is based on the idea the theoretical machine proposed by Alan Turing could perform any function, but not necessarily efficiently. An accelerator pays a lower Turing tariff for its intended functions because operations that are implicit in the module’s circuitry need to be explicitly defined in software when run on a general-purpose processor.

A potential major advantage of moving to accelerator-rich designs in the future is that they do not even have to be confined to using conventional digital logic. The greater emphasis on artificial intelligence (AI) in mainstream computing has encouraged designers to look at alternatives to the CMOS technology used for today’s processors that either perform processing in the

**The term “Turing tariff” is based on the idea that the theoretical machine proposed by Alan Turing could perform any function, but not necessarily efficiently.**

analog domain or use novel switching devices based on electron spin or superconducting techniques to make dramatic energy savings. Though they suffer from poor accuracy and noise, analog and in-memory processors can shrink multipliers that need hundreds or thousands of transistors in the logic domain into just a handful.

Charles Leiserson, professor of computer science and engineering at MIT, says, “There is a lot of really interesting stuff in these approaches that will be helpful for specific, narrow applications. I continue to be impressed by hardware accelerators.”

Users in high-performance computing fields such as machine learning have found accelerators, even with customized code, fail to sustain high throughput when used as part of larger applications. Job startup times and other overheads mean they often leave much of the available performance unused. “The cost-performance ratio is still with the multicores though,” Leiserson adds, because of their relative fungibility and accessibility.

Even with more conventional architectures, communications overheads and the complexity of the memory hierarchy of any multicore implementation can easily trip up developers. “You take out some work from your computation and it slows down, and you say: ‘what?’ If that’s your situation, you can’t architect for that,” Leiserson says. “We need more performance tools and we need hardware to help more there.”

Leiserson and Thompson argue de-

## ACM Member News

### ARTIFICIAL INTELLIGENCE, ROBOTICS, AND INTELLIGENT AGENTS



Maria Gini is a professor in the Department of Computer Science and Engineering at the University

of Minnesota. Her research focuses on artificial intelligence and robotics, with particular interests in robot planning, navigation in unknown environments, coordinated behaviors of autonomous robots, search and rescue applications, and economic agents.

“I tend to be an explorer, and do different things,” Gini notes, adding she has recently started working on conversational agents.

Another area capturing Gini’s interest is swarm robotics, especially with regard to scalability, and how to program them when there are thousands of robots in the swarm. She is interested in distributed systems in which there are multiple robots that are independent but willing to work with each other, rather than operating as adversaries.

Born in Milan, Italy, Gini earned undergraduate and graduate degrees in physics from the University of Milan in 1972. She worked as a Research Associate at the Polytechnic University of Milan (Politecnico di Milano) in Italy, and won a fellowship from the Italian government to study abroad in 1976.

In the U.S., Gini recalls, “I spent time at Stanford University in the AI Lab, that’s where I really learned robotics.”

Gini joined the Department of Computer Science at the University of Minnesota in 1982 as an assistant professor, becoming the department’s first female member. This gender imbalance came as a shock and has led to a lifelong passion for diversity.

“AI can change the world, and there is room for everybody at the table,” Gini remarks.

—John Delaney



velopers should go back to the basics of algorithmic analysis to get better predictability and apply it across entire subsystems. “The great achievement of algorithms is that you can predict coarse behavior by doing a back of the envelope analysis using big-O notation. Even if the constant in front of  $N$  is large,  $N$ -squared is going to be much worse,” Leiserson says.

Researchers see potential improvements in code-generation technologies that understand the underlying hardware and its constraints far better than today and remain portable across target architectures through the use of runtime optimization and scheduling.

Jerónimo Castrillón, chair of compiler construction at Germany’s Dresden Technical University, points to work at that institution into runtime software that can help manage workloads. “You can look at what hardware features you have and percolate them through the stack into the application programming interfaces. For that to work, you need to carry models of the application.”

For example, if an accelerator is unavailable to one module because it is needed by another already running, the scheduler might opt for an alternative compiled for a more general-purpose core instead of holding up the entire application, assuming the compiled code contains enough information to make the analysis possible.

Castrillón believes a shift to domain-specific languages (DSLs) for performance-sensitive parts of the application may be needed, because these can capture more of the developer’s intent. “Usually people think you lose performance if you go to higher levels of abstraction, but it’s not the case if you do the abstractions right.”

Adds Kelly, “With a DSL, the tools can understand that one part is a graph, this other part is a mesh, whereas all a [C or C++] compiler can see is lowered code. Then the compiler is forced to make that uphill struggle to infer what is meant to happen.”

Adaptive heterogeneous systems raise problems of verification and debug: how does the programmer know that a particular implementation still works when it has been re-optimized for a certain fabric at a certain time? One possibility is to use similar formal verifi-

**“Let’s get real about investing in performance engineering. We can’t just leave it to the technologists to give us more performance every year.”**

cation techniques to those employed by hardware designers to check that circuits are functionally equivalent to each other after they have been optimized.

The issue of verification becomes far more difficult when it comes to dealing with accelerators that operate in the analog, rather than the digital, domain, and so do not have the same approach to numerical precision and which will have bounded errors.

AI developers have become accustomed to using loss functions and similar metrics to determine whether neural networks that operate at reduced precision or employ other approximation techniques will perform satisfactorily. Yet there are no methods for doing similar analyses of other types of program, such as physics simulation, where users expect to work with fixed, high-precision formats.

Kelly says more comprehensive numerical analysis will be vital to determining how well an analog accelerator can substitute for a more energy-hungry digital processor. Conventional formal-verification methods, today commonly used in hardware design to check circuit optimizations are correct, do not handle uncertainty. Castrillón says advances in that field, such as probabilistic model checking, may provide a path towards tools that are able to verify the suitability of generated code for an application without demanding bit-level equivalence.

“I don’t know if those things will compose. Or, you can have strong formal analysis on a large system,” Castrillón says.

If composability is not possible, it might fall on programmers to define the levels of accuracy they can tolerate and if a platform cannot meet them, allocate the affected code modules to digital processors that consume more energy or perform the task more slowly.

Although automated code generators may be able to make better use of accelerators than they can do today, there is likely to remain a tension between them and general-purpose cores. Leiserson says while energy concerns push the balance in favor of special-purpose accelerators, generality will likely remain important. “If you have special-purpose hardware, to justify the area it uses, you better be able to use it most of the time.”

If hardware generality continues to prove to be more viable, the main path to energy efficiency and performance in the transition away from the traditional approaches to scaling will be algorithmic in nature, Leiserson concludes. “Let’s get real about investing in performance engineering. We can’t just leave it to the technologists to give us more performance every year. Moore’s Law made it so they didn’t have to worry about that so much, but the wheel is turning.” **□**

#### Further Reading

Wong, H.-S.P., Akarvardar, K., Bokor, J., Hu, C., King-Liu, T.-J., Mitra, S., Plummer, J.D., and Salahuddin, S. **A Density Metric for Semiconductor Technology, *Proceedings of the IEEE*, Vol. 108, No. 4, April 2020**

Hennessy, J.L. and Patterson, D.A. **A New Golden Age for Computer Architecture, *Commun. ACM*, Vol. 62, No. 2 (Feb. 2019)**

Leiserson, C.E., Thompson, N.C., Emer, J.S., Kuszmaul, B.C., Lamson, B.W., Sanchez, D., and Schardl, T.B. **There’s Plenty of Room at the Top: What Will Drive Computer Performance After Moore’s Law? *Science*, 2020 June 5, 368(6495)**

Völz, M. et al. **The Orchestration Stack: The Impossible Task of Designing Software for Unknown Future Post-CMOS Hardware, 1<sup>st</sup> International Workshop on Post-Moore’s Era Supercomputing (2016)**

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

# The State of Virtual Reality Hardware

*Advances in VR hardware could finally take the technology mainstream.*

**F**OR DECADES, VIRTUAL REALITY (VR) has seemed like a futuristic dream that is just around the corner, but never reaches its full potential. This time, however, might really be different. Recent advances in the power of VR hardware, notably the headsets and processors used to produce realistic VR experiences, suggest that VR is finally powerful enough and cheap enough to go mainstream.

VR broadly refers to immersing yourself in a three-dimensional (3D) digital world using sophisticated hardware and software. While a video game is experienced through a screen, VR often is experienced through a headset that shuts out the external world and transports you to a virtual one. It can also be experienced through room-sized systems that use special projectors and glasses to create VR experiences.

Historically, VR has relied on clunky headsets, expensive computers, and complicated peripheral hardware to produce immersive experiences. VR in various forms has been commercially available since the 1990s, but the technology has been widely criticized as too expensive, too complicated, or too imperfect to produce powerful, affordable virtual experiences that inspire consumers to open out their wallets.

That is beginning to change. Today, powerful commercial VR headsets are sold by Sony, Facebook, HTC, and other major technology players. Sophisticated augmented reality (AR) devices (like your smartphone and Google Glass) are available from the likes of Google, Apple, and Microsoft. The market for VR is growing accordingly, with research firm MarketsandMarkets forecasting industry growth to reach \$20.9 billion in 2025, from \$6.1 billion in 2020.



Why is VR (finally) having its day in the sun?

It all comes down to better hardware. VR heavyweights now are able to produce headsets that are cheaper and more powerful than models from just a few years ago. As a result, consumer demand for headsets is rising, driving more innovation and investment in VR hardware. Companies are even researching entirely new techniques and designs to make the next generation of VR hardware so light and powerful that it transforms one's daily life.

"Over time, we would like a device that is nearly the size of your reading glasses or sunglasses, but performs all of the functions of your smartphone, tablet, PC, and even TV, and enables new, 3D (three-dimensional) and spatial functions," says Siddharth Saxena, founder and CEO of Oblix VR, a VR software startup.

That day might not be far off.

## Better Hardware, Better VR

The top three players in VR headsets by sales are Sony, Facebook, and HTC, according to research provided by Statista. In 2019, 5.7 million VR headsets were shipped, according to research from SuperData.

Back in 2016, commercial VR sys-

tems required users to connect a headset, controllers, and sensors to an external high-end computer, says Saxena. "This was an expensive, bulky, and inconvenient setup," he says.

Today, however, systems like Facebook's Oculus Quest 2 are all-in-one VR platforms with built-in processors that require no external computer at all. This represents a huge leap forward from just a few years ago, says Bill Myers, director of Emerging Technology at S3 Technologies.

In 2017, Myers started a VR arcade that offered consumers access to sophisticated VR stations. Each station used an HTC Vive headset and a high-end computer, which cost a total of \$2,500 at the time. Now, he says, consumers can experience the same level of VR immersion with a headset like those available today for just a few hundred dollars.

Why the big jump forward?

In Myers' view, VR has hit a consumer tipping point. By 2016, the technology finally was good enough and cheap enough (though still limited to those with deep pockets) that consumer demand drove continued progress. Companies were incentivized to manufacture more VR headsets. Suppliers improved their tooling and lensing capabilities to serve demand, and better processing power allowed developers to create truly immersive experiences.

"Now, when a developer sits down to create an experience, it doesn't have to be what the previous generation was, which was a lot of low-polygon graphics," says Myers. This is leading to higher levels of performance expected from consumer headsets, and growing consumer demand.

"With advances in inside-out tracking and improved chipsets like the Qualcomm XR2 [the chip used in the

Quest 2], the standard for simple, portable VR is now a lot higher,” says Angel Say, CEO and cofounder of Resolve, a company that uses VR to help construction companies review building designs faster and more economically. “Receiving a VR headset as a gift used to be a burden if you didn’t have the right-specification PC or knowledge to set it up; now, it’s as easy as unboxing and setting up a phone for the first time.”

This was not the case when you had to install desktop software, install drivers, calibrate sensors, and deal with bulky peripherals, says Say. “Stand-alone VR headsets are enabling anyone to pick up a VR headset and, within 15 minutes, be up and running.”

While hardware has come a long way in a few short years, true mass adoption will only come by further reducing the size of devices while maintaining or improving capabilities, according to Saxena. “I don’t think we’re quite where we need to be for mass adoption, but we’re close,” he says.

Research teams are trying to get VR over the finish line.

### The Next Generation

Despite advancements in processing power and price, today’s VR hardware is still limited. Sony’s Playstation VR headset, a popular model, weighs more than 1.3 pounds. The heaviest of the most popular headsets, the Valve Index, weighs almost two pounds. Researchers are trying to get around weight constraints with designs that use innovative techniques and materials.

In 2020, Facebook Reality Labs researchers Andrew Maimone and Junren Wang released a paper outlining how holographic optics could be used to create ultra-lightweight VR headsets.

In the paper, titled “Holographic Optics for Thin and Lightweight Virtual Reality,” Maimone and Wang outline the problem facing today’s headsets. Commercially available headsets use “curved optics of solid glass or plastic, which has limited designs to goggles-like form factors.” While goggle-like VR headsets have become lighter and slimmer, they’re still relatively heavy and bulky. The researchers suggest using a combination of

new optical design techniques to overcome the problem.

One such technique is polarization-based optical folding, a way to design lenses so light bounces in the right way to the human eye so on-screen images are displayed properly—but the light doesn’t need to physically travel as far as it does in traditional optics. That makes the space needed for VR optics smaller.

The other technique under consideration is holographic optics, an optics technology that “bends light like a lens but looks like a thin, transparent sticker,” according to Facebook’s summary of the research. Holographic optics replace glass or plastic lenses, making the resulting VR headset much lighter.

In fact, these advances could make the VR headsets of tomorrow, with proposed designs less than 10mm in thickness. While still in the prototype phase, the research suggests a possible approach to VR that almost entirely eliminates the need for bulky hardware.

That means the future of VR hardware could actually be closer to that of augmented reality (AR), experts say. With the ability to project immersive imagery through thin, lightweight lenses, the lines between the virtual world and the real one could get extremely blurry.

The new proposed technology from Facebook actually controls the light within a thin lens. That can theoretically solve the issue of external light interfering with simulated images projected to the user, which could further reduce barriers to integrating VR and AR. This is important because, as Facebook chief scientist Michael Abrash explains, “Even when AR is something that everybody uses, 99% of the photons that hit your eyes or your retinas are actually still going to come from the real world.”

“We’re already seeing glimpses of VR devices converging with AR,” says Say. “Being able to toggle between being completely immersed in a digital world and overlaying digital info on the real world is really powerful.”

It’s a future that VR hardware leaders are actively exploring.

“Facebook is in a unique position from a hardware standpoint because

they are producing both AR and VR headsets simultaneously,” says Myers. He points to the company’s partnership with Ray-Ban to release a pair of smart glasses. The product won’t have AR features to start, but is seen as the first step toward Facebook’s Project Aria (announced at the same time), an initiative to build true wearable AR devices.

If these plans come to fruition, expect VR, AR, or a combination of these technologies to integrate more seamlessly into your everyday life.

“These products will be lifestyle products,” says Myers. “In order to get our tasks done in the next five years, we’ll have access to these pieces of technology that allow us to work and play in a whole new way.” □

### Further Reading

Heath, A.

Facebook’s Chief Scientist: Mass Adoption of AR Is Years Away, *The Information*, Jan. 2020, <https://bit.ly/3ozSnQB>

Maimone, A. and Wang, J.

Holographic Optics for Thin and Lightweight Virtual Reality, *Facebook Reality Labs*, July 2020, <https://bit.ly/37NuhLk>

Virtual Reality Market with COVID-19 Impact Analysis by Offering (Hardware and Software), Technology, Device Type (Head-Mounted Display, Gesture-Tracking Device), Application (Consumer, Commercial, Enterprise, Healthcare) and Geography, *MarketsandMarkets*, Aug. 2020, <https://bit.ly/3qDtaqf>

Virtual Reality Market Size, Share & Trends Analysis Report By Device (HMD, GTD), By Technology (Semi & Fully Immersive, Non-immersive), By Component, By Application, By Region, And Segment Forecasts, 2020–2027, *Grand View Research*, Jun. 2020, <https://bit.ly/33Su88e>

Unit shipments of virtual reality (VR) devices worldwide from 2017 to 2019 (in millions), by vendor, *Statista*, Dec. 2018, <https://bit.ly/39Uy4JI>

Holographic optics for thin and lightweight virtual reality, *Facebook Research*, Jun. 29, 2020, <https://research.fb.com/blog/2020/06/holographic-optics-for-thin-and-lightweight-virtual-reality/>

2019 Year in Review, *SuperData*, 2020, <https://www.superdataresearch.com/reports/p/2019-year-in-review>

Logan Kugler is a freelance technology writer based in Tampa, FL, USA. He has written for over 60 major publications.



# Technological Responses to COVID-19

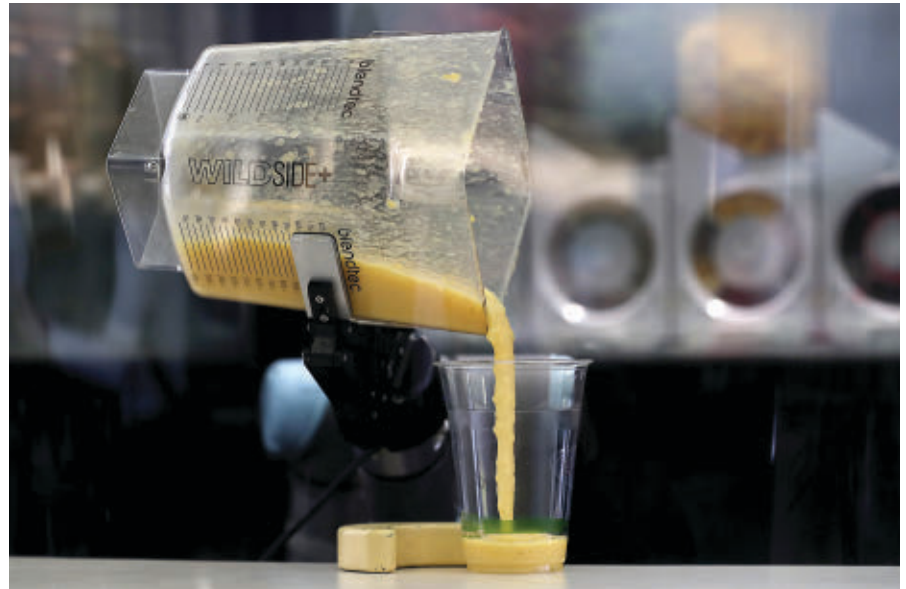
*Companies are finding new ways to enforce social distancing, clean public spaces, and provide substitutes for human workers.*

**T**HE IMPACTS OF the COVID-19 pandemic are likely to be felt for years to come, regardless of the presence and availability of a vaccine. Physical measures adopted by humans, such as social distancing or wearing masks, are likely to be utilized for years to come, along with technological developments deployed in both public and private spaces that are focused on enforcing social distancing, enabling more efficient cleaning and disinfecting of spaces, and driving more automation and intelligence to reduce humans' direct physical interaction with each other.

Some companies and individuals feel the best way to avoid COVID-19 or other viruses is to simply avoid all unnecessary human contact. As such, many companies have introduced or fast-tracked the use of automation to lessen their reliance on human workers, as well as to enhance their responsiveness to customer queries.

For example, beginning last fall, the White Castle burger chain planned to test Flippy, a robot arm that can cook French fries and other foods. Made by Miso Robotics, Flippy can free up employees for other tasks, like disinfecting tables or addressing delivery orders, while reinforcing a touch-free environment during food preparation, important to people concerned about the spread of germs. Miso says Flippy currently costs \$30,000, plus a \$1,500 monthly service fee, but the company expects to shift to a different business model by the middle of this year. This new model charges users a higher monthly service fee in lieu of an up-front charge for the robot.

Other companies also have rolled out food-preparation robots during the pandemic, such as Hayward, CA-based Chowbotics, which has deployed Sally, a robot about the size of a refrigerator that can make up to 65 bowls of sal-



**A robot at the Blendid kiosk on the University of San Francisco campus pours a smoothie it just made into a cup. The robot, which takes the place of a human employee, is capable of making up to 45 smoothies per hour.**

ads before needing to be refilled. The company says it has been deployed at grocery stores, hospitals, and college campuses, such as Big Y Supermarket, University of Arkansas for Medical Sciences, and Elmira College, among others. Similarly, Blendid, a Silicon Valley startup, sells a robotic kiosk that can make fresh smoothies without human intervention, guided by a smartphone app that allows users to customize their drinks. These kiosks are currently in use at the University of San Francisco Market Café, at Charlie Brown's Café at Sonoma State University, and at Plug and Play, a tech center in Sunnyvale, CA.

Despite the increasing use of robots and automation, avoiding all human contact is impossible, particularly in the retail environment. Retail spaces—and other public spaces—are among the most challenging environments to keep clean under normal circumstances, given the large number and variety of people and objects within the space,

the extensive surface area on which the virus could settle, and the tendency of humans to touch objects indiscriminately. Two primary approaches are being utilized to help keep customers safe.

The first is by implementing technology to help address social distancing measures, by tracking not only how many people are inside a store at a given point of time, but also to help ensure they are not bunching up or crowding together. For example, Pune, India-based technology company Glimpse Analytics refocused its artificial intelligence-based analytics device to help alert retail stores or offices to violations of occupancy limits, or situations where social distancing or personal protective equipment (PPE) mandates such as mask wearing are not being observed.

According to Kakshil Shah, one of Glimpse Analytics co-founders, the technology uses a store's existing CCTV cameras to capture images that are analyzed on "edge," or local, devices, using

machine learning to assess a number of demographic or behavioral indicators about shoppers, including the number of shoppers within the store at any given time, the size and amount of time spent waiting in line, in-store traffic patterns and heat maps, as well as basic demographic data such as ethnicity, age, and gender. Because the data is processed by Glimpse Analytics on a standalone edge device within the store, no data is sent back to a central processor offsite, or stored in perpetuity by the vendor.

“There are two primary features that are actively used,” Shah says. “One is tracking how many people are actually inside of a store or a mall. The other thing we are seeing being used is for [monitoring] queues,” either within stores or in malls, which helps provide insight to shopping center operators who want to see where people are bunching together, and identify which stores or areas within a store are most popular. Shah also notes that some retailers are using the computer vision technology to determine whether customers are wearing a face covering, and if they are wearing it properly.

The purpose of the system is to capture data on how and where people are congregating, or whether they are not adhering to health protocols. Rather than being used to call out individuals for non-compliance, the data is extracted so the store can implement policy or tactical changes to improve social distancing or address staffing issues to help enforce mask wearing. Glimpse says it is working with a number of brands in In-

dia (Future Group, Samarth Mart, and The Souled Store), in Kuwait (Synergy United Co.), in London (at several malls and the London Underground,), and in a business park in the U.S.

Still, many people are still wary of entering indoor spaces, particularly elderly and immune-compromised individuals. As such, many companies have expanded curbside delivery and pickup options, which are designed to allow people to avoid crowds and lines in the stores, as well as limiting their interactions with store staff. The challenge for many retailers lies in managing a curbside pickup program, particularly if they share curbsides and parking spaces with other retailers that also may be offering curbside pickup.

RE Insight, an Irvine, CA-based technology company that provides hardware and software solutions to retail store owners and operators such as shopping malls and shopping centers, offers vQueue+, a hardware and software platform that allows properties to track and analyze visitor counts, manage and encourage social distancing protocols, and communicate with shopping center visitors to manage reservations, pick-up orders, or handle virtual queuing quickly and efficiently. This text-based solution is designed to interface with each individual store’s customer management solution, allowing visitors to create reservations for merchandise pickup, monitor their position in a queue, and receive update notifications on a mobile device via text.

Within a shopping center or mall,

all of the disparate retailers are able to offer curbside pickup, and have it coordinated and managed through a single, common platform, reducing competition and confusion among retailers while ensuring better customer service. Most importantly, customers can book pickup reservations at multiple stores within the shopping center, and the platform will route them to the appropriate pickup area, taking into account pick-up queues and traffic.

“The challenge operators have is that people want to pick up at different stores,” says Quinn Munton, president and CEO of RE Insight. “Lowe’s is doing it differently than Best Buy, and Ulta [Beauty] wants to do it differently than PetSmart. They all want the prime parking spots blocked off, they want to use their own signage, and they don’t have any way to communicate with [customers.] And so, we’ve seen a lot of early disasters with owner-operators where they rolled something out and then it failed because they didn’t have a communication tool. So we built a platform that will integrate with apps, and will integrate with their Web portal so that they can make it a seamless experience from beginning to end.”

As life continues to return to normal, retailers and other spaces that are used by the public have been forced by COVID-19 to deploy more frequent and robust cleaning measures. One way to handle this task more efficiently than hand-wiping every item, shelf, or surface in the space is by using automated robots to spray disinfectant that can

## Milestones

# 2020 ACM Gordon Bell Prize Awarded for Tool Simulating Interactions of 100 Million Atoms

ACM named a nine-member team from Chinese and American institutions to receive the 2020 ACM Gordon Bell Prize for their project, “Pushing the limit of molecular dynamics with *ab initio* accuracy to 100 million atoms with machine learning.”

Winning members of the research team include Weile Jia, University of California, Berkeley; Han Wang, Institute of Applied Physics and Computational Mathematics, Beijing, China; Mohan Chen, Peking University; Denghui Lu, Peking University;

Lin Lin, University of California, Berkeley and Lawrence Berkeley National Laboratory; Roberto Car, Princeton University; Weinan E, Princeton University; and Linfeng Zhang, Princeton University.

Molecular dynamics is a computer simulation method that analyzes how atoms and molecules move and interact. Simulations of molecular dynamics allow scientists to gain a better sense of how a system progresses over time.

In their winning paper, the

team introduced Deep Potential Molecular Dynamics (DPMD), a machine learning-based protocol that can simulate a more than 1-nanosecond-long trajectory of over 100 million atoms per day. The team wrote, “The great accomplishment of this work is that it opens the door to simulating unprecedented size and time scales with *ab initio* accuracy. It also poses new challenges to the next-generation supercomputer for a better integration of machine learning and physical modeling.”

The ACM Gordon Bell Prize tracks the progress of parallel computing and rewards innovation in applying high-performance computing to challenges in science, engineering, and large-scale data analytics. The award was presented by ACM President Gabriele Kotsis and Bronis de Supinski, chair of the 2020 Gordon Bell Prize Award Committee, during the virtual International Conference for High Performance Computing, Networking, Storage and Analysis (SC20).

kill viruses or other pathogens. An example of the technology comes from Pratt Miller Mobility (PMM), which has deployed its Autonomous Disinfecting (LAAD) vehicle at the Gerald R. Ford International Airport in Grand Rapids, MI. As part of a grant from the Michigan Economic Development Corporation (MEDC) and PlanetM, the state's mobility initiative, the LAAD was deployed in July 2020 to demonstrate how large public areas could be efficiently and reliably disinfected.

Using a combination of computer vision technology and LiDAR (light detection and ranging), the connected, autonomous LAAD vehicle can navigate autonomously throughout public spaces, delivering a measured amount of FDA-approved disinfectant through a multi-head electrostatic sprayer array far more efficiently and reliably than humans could. Through the use of sensors and data analysis, the autonomous platform monitors and guarantees delivery of the disinfectant and, depending upon the size of tank used to store disinfectant, can cover areas the size of airport terminals, arenas, or shopping malls.

"In its current configuration, LAAD holds seven gallons of solution, which is ideal for long runs like airport terminals," says Chris Andrews, Pratt Miller Director of Mobility and Innovation. "It not only alleviates having to refill the tank, but it monitors, documents, and reports on what was covered."

Andrews also highlights the LAAD's flexibility. "The modularity of the platform allows users to customize the system for their individual needs," he says. "If you need more or less solution and coverage, we can configure the platform for the application."


Andrews said the LAAD would be further evaluated and refined, and expected to make deployment announcements late last year, anticipating rollouts to big-box retailers, sports stadiums and arenas, and shopping malls.

Other disinfecting solutions are focused on using UV-C, a type of ultraviolet light, on a more targeted basis, particularly when dealing with objects likely to be touched by the public, such as menus, keypads, pens, or other small items. Vioguard's Cubby Plus uses UV-C light contained within a drawer-like case that can be used to target coronavirus, along with a host of other bacteria and

## Other disinfecting solutions focus on using UV-C, a type of ultraviolet light, on a more targeted basis, particularly when dealing with objects likely to be touched by the public.

viruses. The level of UV-C light required to kill viruses and bacteria is potentially harmful to humans if they look directly at the source, so a contained solution works better for high-traffic areas, such as retail and restaurants.

The Cubby Plus cabinet solution is particularly useful for items that are frequently handled by staff and customers, such as credit card holders at restaurants, pens, and even small merchandise items such as jewelry that need to be disinfected regularly.

"I sat down at a restaurant and the waiter came and put a laminated menu on the table," recalls Mark Beeston, VP of sales and marketing for Vioguard. "And I thought that, 'you know, are these menus getting cleaned or are they just being collected and then stacked back in the back?' So a laminated menu or a paper menu could be run through the Cubby Plus and disinfected." 

### Further Reading

*Cadnum, Jennifer, et al.*  
Evaluation of an electrostatic spray disinfectant technology for rapid decontamination of portable equipment and large open areas in the era of SARS-CoV-2. *Am J Infect Control.* 2020 Aug; 48(8): 951–954. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7275188/>

Large Area Autonomous Disinfecting Vehicle information site, Pratt & Miller, <https://www.prattmiller.com/laad>

Miso Robotics - Meet Flippy, <https://www.youtube.com/watch?v=yLO-OgRpobo>

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

© 2021 ACM 0001-0782/21/2 \$15.00

### Milestones

## Team Receives 1<sup>st</sup> ACM Gordon Bell Special Prize for High-Performance Computing-Based COVID-19 Research

The first 2020 ACM Gordon Bell Special Prize for High Performance Computing-Based COVID-19 Research was presented to a 12-member team for their project "AI-Driven Multiscale Simulations Illuminate Mechanisms of SARS-CoV-2 Spike Dynamics." The Prize was awarded in recognition of outstanding research achievement toward understanding the COVID-19 pandemic through the use of high-performance computing (HPC).

In their paper, the winning team develops a generalizable artificial intelligence-driven workflow leveraging heterogeneous HPC resources to explore the time-dependent dynamics of molecular systems. They used the workflow to investigate mechanisms of infectivity of the SARS-CoV-2 spike protein, the main infection machinery of the virus.

Their workflow enables more efficient investigation of spike dynamics in a variety of complex environments, including within a complete SARS-CoV-2 viral envelope simulation that contains 305 million atoms and shows strong scaling on Oakridge National Laboratory's Summit supercomputer using nanoscale molecular dynamics (NAMD) software.

The team had several novel scientific discoveries, including elucidation of the spike's full glycan shield, the role of spike glycans in modulating the infectivity of the virus, and the characterization of flexible interactions between the spike and the human ACE2 receptor. They also demonstrated how AI can accelerate conformational sampling across different systems and pave the way for the future application of such methods to additional studies in SARS-CoV-2 and other molecular systems.

A cash prize in the amount of \$10,000 accompanies the award, which was conceived and funded by Gordon Bell, a pioneer in high-performance computing and researcher emeritus at Microsoft Research.





DOI:10.1145/3442371

Yannis Bakos, Hanna Halaburda, and Christoph Mueller-Bloch

► Marshall Van Alstyne, Column Editor

# Economic and Business Dimensions When Permissioned Blockchains Deliver More Decentralization Than Permissionless

*Considerations for the governance of distributed systems.*

**P**ERMISSIONLESS BLOCKCHAIN SYSTEMS inspired by Bitcoin and related crypto-ecosystems are frequently promoted as the enablers of an open, distributed, and decentralized ideal. They are hailed as a solution that can “democratize” the economy by creating a technological imperative favoring open, distributed, and decentralized systems, platforms, and markets. We argue that such claims and expectations, while they may be fulfilled under certain circumstances, are frequently exaggerated or even misguided. They illustrate a tendency to equate open access with decentralized control in distributed architectures, an association that while possible is far from guaranteed. When enterprise, social and economic activities are “put on the

blockchain” in order to avoid centralized control, permissioned governance may offer a more decentralized and more predictable outcome than open permissionless governance offers in practice.

## Access and Control in Distributed Systems

Information systems can be characterized on three key dimensions: architecture, which can be concentrated or distributed,<sup>17</sup> access, which can be permissionless or permissioned,<sup>1</sup> and control (that is, the locus of decision rights), which can be centralized or decentralized.<sup>7</sup> These dimensions are not binary, and the associated labels should be thought as endpoints of a continuum.

Permissionless systems do not restrict who has access, and thus are also referred

to as open-access.<sup>a</sup> For instance, in principle anyone can post source code on GitHub, edit a Wikipedia article, or validate bitcoin transactions. Permissioned systems only grant access to qualified users. The distinction for control focuses on who gets to make decisions. Centralization implies that decisions are made by a single person or a small group; decentralization means that decision rights are widely distributed.<sup>7</sup>

It has long been argued that concentrated architectures favor permissioned access and centralized control because these types of access and control reinforce

<sup>a</sup> We will often use the term open-access for permissionless systems to avoid any confusion from repeated use of the terms permissionless and permissioned.



the benefits of these architectures;<sup>7</sup> see for instance early arguments about Grosch's law for computer hardware,<sup>6</sup> or the administration of early databases. However, as technology evolved to enable or even favor distributed system architectures, open access and decentralized control emerged as feasible alternatives.

In this column, we examine the issues of open vs. permissioned access and centralized vs. decentralized control in distributed systems, focusing on blockchain implementations. We argue that while distributed architectures may enable open access and decentralized control, they do not preordain these outcomes. Furthermore, while open access and decentralization are frequently thought as complementary,<sup>14</sup> experience from real-world applications suggests that the opposite can also be true: open access may result in essentially centralized control, while permissioned systems may be able to better support decentralized control.

### How Permissioned Systems Can Be More Decentralized

While this possibility may seem counterintuitive at first, it can be understood as a consequence of the need to provide

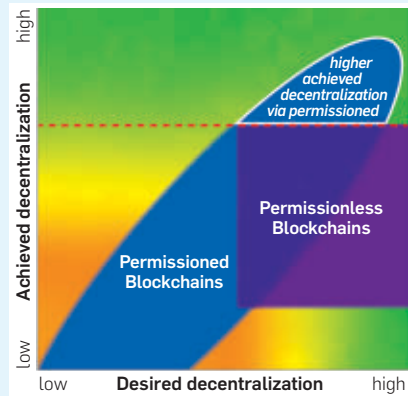
appropriate incentives to system participants, especially the ones that operate the technology after its implementation. The economic theory of Incomplete Contracts<sup>10,11</sup> shows that when an agent's actions affect the value of an asset, such as an information system, but these actions cannot be contractually specified (for example, because the necessary behaviors cannot be adequately verified), the agent should be given corresponding control or ownership to maximize agent incentives. Van Alstyne, Brynjolfsson, and Madnick<sup>18</sup> apply this argument to derive design principles for databases; for instance, when maintenance of data quality is important, any independent local data partitions should be locally controlled.

These considerations apply to systems beyond databases, however. In the blockchain context certain system participants can be indispensable in the sense that the system's operation and value generation will depend on actions that cannot be contractually specified. In such cases, the need to incentivize these participants will likely lead to outcomes where they effectively control the parts of the system over which they are indispensable. Depending on the particular situation, this can lead

towards either centralized or decentralized control. For instance, in an open access and fully distributed environment it may be infeasible to incentivize participants to adequately provide functions like quality control or coordination of system development and evolution. To address this problem, centralized solutions emerge de facto, such as the hierarchy of the small number of developers controlling open source projects,<sup>5</sup> or the hierarchy of editors in Wikipedia.<sup>16</sup> This is because expertise, reputation, time, or money can all be required to take advantage of open access and decentralized control. The higher these costs are, the fewer the people that want to participate, which contributes to this centralization in practice.<sup>9</sup>

It is thus important to distinguish between how governance is envisioned and how it is enacted. Without this distinction, the potential for decentralization in open-access systems is often overstated, while the potential of permissioned systems in achieving decentralization is not fully recognized. Open-access systems in principle allow for arbitrary decentralization, but cannot guarantee decentralization at any level, as the actual level of decentralization is the result of individual decisions. This

### Decentralization in permissioned and permissionless blockchains.



ambiguity of outcome is important when open access and decentralization are desirable or even the reason technologies like blockchain are adopted, for instance when there is a goal to promote “democratization,” to avoid intermediaries that are in a favorable position to extract economic rents, or when there are no parties that can be trusted with regulating permissioned access or making decisions for the majority of users.

### The Case of Blockchain

Blockchain technology provides a prominent illustration: While blockchain systems are distributed architecturally, control can be centralized and/or access can be permissioned. Permissionless blockchains such as Bitcoin’s do not restrict who can validate transactions. Permissioned blockchains, however, only grant these rights to selected agents.<sup>3</sup> With the growing interest in permissioned blockchains, it is crucial to understand whether these blockchains can actually deliver on the promise of decentralization.

The Bitcoin ideal<sup>15</sup> has created the expectation for blockchain technology to universally deliver open, decentralized, “democratic” systems that bypass controlling intermediaries. Real-world applications of blockchain systems, however, show that this ideal is the exception rather than the rule.<sup>8</sup> While permissionless blockchains like Bitcoin do not restrict who can validate transactions, and thus can allow access close to the permissionless ideal, often control is far from decentralized. In the absence of formal checks for the underlying centralization forces, centralization emerges in practice, for instance exercised by large emergent min-

ing pools with de facto operational power.<sup>2</sup> This means that the promise of blockchain to remove trusted third parties remains unfulfilled. For example, in May 2018 alone, five open-access blockchains were compromised due to overt centralization.<sup>12</sup>

Permissioned blockchains have been criticized for not being truly decentralized (for example, Beedham<sup>4</sup>) in contrast to open-access blockchains. This is because they restrict who can become a validator, which is decided by a gatekeeper giving permissions. In Libra, a cryptocurrency spearheaded by Facebook, gatekeeping is the task of the Libra Association, which is governed by a council of all existing validator nodes. Therefore, the existing validator nodes jointly serve as a gatekeeper and decide whether a new validator is allowed to join the network.<sup>13</sup> The gatekeeper can often also encourage participation through off-blockchain channels.

### Designing for Decentralization

While not fully decentralized by design, the governance structure of permissioned systems can guarantee a certain level of decentralization. For instance, consensus mechanisms for permissioned blockchains can be designed in a way that guarantees a large number of nodes get a say in the validation process. Moreover, a large number of validators can be guaranteed through off-blockchain negotiation, enforcing their participation. In open-access blockchains however, this is impossible to guarantee—decentralization (or indeed, centralization) can only emerge as a potential outcome of free individual decisions.

Creating a permissioned blockchain that offers more decentralization than an open-access blockchain requires careful design. For instance, the power to grant and especially to revoke validation rights is central, and thus in order to promote decentralization in permissioned blockchains it is necessary to decentralize the gatekeeping function. If a central gatekeeper can arbitrarily revoke validation rights, it could easily take over and centralize the entire blockchain. While it is possible to guarantee a certain degree of decentralization, it is crucial to get the blockchain governance right.

### Conclusion

The case of blockchain technology highlights an important consideration for

the governance of distributed systems. System designers must account for the interactions between access and control, and make design choices based on their goals. As illustrated in the figure, if the primary objective for a distributed system is decentralization, a well-designed permissioned system may be better positioned to achieve it in practice. □

### References

1. Abadi, M. et al. A calculus for access control in distributed systems. *ACM Transactions on Programming Languages and Systems (TOPLAS)* 15, 4 (1993), 706–734.
2. Arnosti, N. and Weinberg, S.M. Bitcoin: A natural oligopoly. (2018); arXiv preprint arXiv:1811.08572.
3. Beck, R., Müller-Bloch, C., and King, J.L. Governance in the blockchain economy: A framework and research agenda. *Journal of the Association for Information Systems* 19, 10 (Oct. 2018), 1020–1034.
4. Beedham, M. Here’s the difference between ‘permissioned’ and ‘permissionless’ blockchains. *The Next Web* (Nov. 5, 2018); <https://bit.ly/37aoL6E>
5. Crowston, K. and Howison, J. The social structure of free and open source software development. *First Monday* (2005).
6. Grosch, H.R. High speed arithmetic: The digital computer as a research tool. *Journal of the Optical Society of America* 43, 4 (1953), 306–310.
7. King, J.L. Centralized versus decentralized computing: organizational considerations and management options. *ACM Computing Surveys (CSUR)* 15, 4 (1983), 319–349.
8. Halaburda, H. Blockchain revolution without the blockchain? *Commun. ACM* 61, 7 (2018), 27–29.
9. Halaburda, H. and Müller-Bloch, C. Will we realize Blockchain’s promise of decentralization? *Harvard Business Review* (Sept. 2019).
10. Hart, O. Incomplete Contracts and Control. Prize Lecture for the Nobel Memorial Prize in Economic Sciences, Stockholm, (Dec. 8, 2016).
11. Hart, O. and Moore, J. Property rights and the nature of the firm. *Journal of Political Economy* 98, 6 (1990).
12. Hertig, A. Blockchain’s once-feared 51% attack is now becoming regular. *CoinDesk* (June 8, 2018), <https://bit.ly/2KiaQ5a>
13. Libra. How to Become a Founding Member. *Libra* (Jan. 21, 2020); <https://bit.ly/3a4sLYa>
14. Liu, M., Wu, K., and Xu, J.J. How will Blockchain technology impact auditing and accounting: Permissionless versus permissioned Blockchain. *Current Issues in Auditing* 13, 2 (2019), A19–A29.
15. Nakamoto, S. *Bitcoin: A Peer-to-Peer Electronic Cash System*. 2008.
16. Ortega, F., Gonzalez-Barahona, J.M., and Robles, G. On the inequality of contributions to Wikipedia. In *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008) (2008)*. IEEE, 2008, 304–304.
17. Tanenbaum, A.S. and Van Steen, M. *Distributed Systems: Principles and Paradigms*. Prentice-Hall, 2007.
18. Van Alstyne, M., Brynjolfsson, E., and Madnick, S. Why not one big database? Principles for data ownership. *Decision Support Systems* 15, 4 (1995), 267–284.

**Yannis Bakos** (bakos@stern.nyu.edu) is an Associate Professor of Information Systems at NYU Stern School of Business, New York University, New York, NY, USA.

**Hanna Halaburda** (hhalaburda@gmail.com) is an Associate Professor of Information Systems at NYU Stern School of Business at New York University, New York, NY, USA.

**Christoph Mueller-Bloch** (chmy@itu.dk) is a Postdoctoral Researcher at the IT University of Copenhagen, Denmark.

The authors would like to thank *Communications* section editor Marshall Van Alstyne, the rest of the editorial team, and the two anonymous referees for their helpful comments and suggestions.

Copyright held by authors.



▶ Mark Guzdial, Column Editor

## Education

# CAPE: A Framework for Assessing Equity throughout the Computer Science Education Ecosystem

*Examining both the leading indicators of equity in CS and the lagging indicators of student outcomes.*

**W**OMEN AND PEOPLE of color are underrepresented in the U.S. computing workforce<sup>5,6,8</sup> and in computing majors and coursework in higher education and K-12.<sup>1</sup> Addressing this lack of diversity requires interventions in both the culture and practice of the computing industry as well as earlier in the education pipeline. The National Science Foundation has made significant investment over the past decade to broaden participation in computing (BPC) through programs such as CS10K, RPP for CS, and BPC Alliances like Expanding Pathways in Computing (ECEP). The release in 2017 of a new high school course in the U.S. called, AP CS Principles, has resulted in some improvements in diverse participation in high school CS,<sup>3</sup> and the Computing Research Association has reported modest improvements in the enrollment of women and students of color in introductory CS major courses.<sup>2</sup> However, it remains to be seen whether this limited progress will result in substantive improvements to diversity in the computing industry.

Moving the needle on diverse representation in computing coursework



is often the de facto, end-of-the-line measure of success in these various efforts. Less attention has been paid, however, to the entire ecosystem of CS education and the precursors or root causes of underrepresentation. The CAPE framework is a lens for assessing equity not simply as an end product, but as an integral component to each element of the systems that support computing education. The frame-

work addresses four key components of CS education: *Capacity for*, *Access to*, *Participation in*, and *Experience of* equitable CS education (CAPE). The CAPE pyramid shown in the figure in this column is meant to illustrate how the four components of the framework interact progressively, building and relying on the previous component. For example, if students are to have equitable experiences learning CS,

they must first participate in CS courses and programs. If students are to choose to participate in CS, they must first have equitable access to CS courses and programs. If schools and universities are to provide students access to CS, they must first have the capacity to offer inclusive CS instruction for all students, not just a privileged few. We posit that until we begin to address the root causes of underrepresentation in CS at each of these levels, the U.S. will continue to struggle in developing a CS education system and workforce that fully leverages the contributions of our diverse national populace.

Equity research often examines disparities in student outcomes such as Advanced Placement (AP) CS passing rates or degree completion. But these types of disparities are *lagging* indicators of inequity and focusing solely on such metrics ignores the varied systemic barriers to equitable outcomes that were put in place long before students enrolled in courses or completed a degree. The CAPE framework can help instructors, researchers, practitioners, and policymakers to examine the ecosystem in which K–16 CS education is embedded and create a deeper understanding of the precursor conditions and *leading* indicators of systemic inequities in the experience of CS for historically underrepresented populations, including women, students of color, students from families with limited financial resources, students with disabilities, and students who live in rural communities. Each of the four levels of the framework carries important implications for how we think about, measure, and ultimately impact equity in CS education.

### Capacity for CS Education

Capacity for CS education refers to the availability of resources to support and maintain high-quality CS instruction. These resources may include faculty, funding, and policies that make implementing CS instruction possible and inclusive. At this level, researchers can examine equity through questions such as:

- ▶ High School

- ▶ Are there differences based on student socioeconomic status in the proportion of schools that employ certified CS teachers?

## The CAPE framework is a lens for assessing equity not simply as an end product, but as an integral component to each element of the systems that support computing education.

- ▶ College

- ▶ How does a shortage of CS faculty impact opportunity for underrepresented students to major in CS?

- ▶ How does faculty capacity impact policies around access to CS coursework for non-majors?

Each of these questions around capacity to deliver CS instruction has implications for the eventual equitable access to and participation in CS both in K–12 and in higher education. For example, if trained and certified teachers are disproportionately employed by wealthier school districts, the capacity of schools serving primarily low-income students to provide high quality CS courses will be severely constrained. Regarding undergraduate computing, if an institution is struggling to serve all students who hope to major in CS due to a shortage of faculty, are admission filters in place that only accept students with prior experience

in CS into the major? If so, how does that impact students from low-income households who are less likely to attend high schools that offer CS or have access to CS experiences outside of school? How do these capacity issues exacerbate the challenges of diversifying undergraduate CS opportunities? Each of these questions about capacity address issues that can impact, at a very early stage, whether traditionally marginalized students have opportunities to engage in CS education.

### Access to CS Education

At the high school level, access can be operationalized as attending a school that offers CS courses. At the undergraduate level, access can address both access to CS as a CS major as well as access to CS coursework for non-majors. Equity in access to CS can be explored by examining questions such as:

- ▶ High School

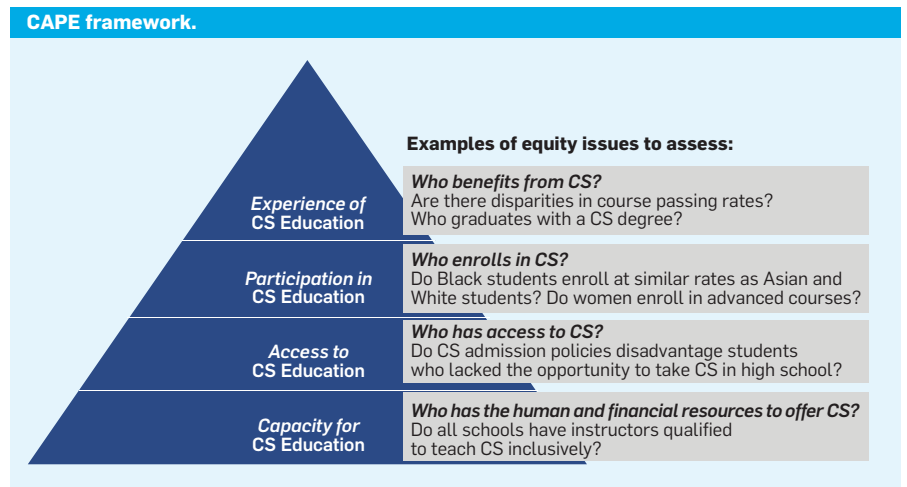
- ▶ How do rural and urban/suburban schools differ in terms of offering CS courses?

- ▶ College

- ▶ What are the barriers and facilitators for community college transfers into undergraduate CS majors at four-year institutions? How do these barriers/facilitators impact diversity at four-year institutions?

- ▶ How do majors in other STEM fields access CS courses? How does access to CS coursework for non-majors impact learning opportunities in CS for males and females differentially?

As of 2020, only 47% of U.S. high schools offered a single computer science course. Moreover, this limited access is not equitably distributed across



diverse populations.<sup>1</sup> Even when CS courses are offered, access to multiple courses or more advanced CS coursework is often highly correlated with affluence. How does this lack of access ultimately impact diversity in CS majors in college and industry?

With respect to undergraduates, low-income students and students of color are disproportionately enrolled in community colleges (as opposed to four-year universities). In the U.S., 31% of undergraduate students from families with the lowest income (lowest quartile) enrolled in community colleges compared to only 17% of undergraduates in the highest income quartile.<sup>7</sup> Similarly, 41% of Black undergraduates and 48% of Hispanic/Latino undergraduates enrolled in community colleges first compared to 34% of White undergraduate students.<sup>4</sup> Because of this, policies at four-year institutions that effectively prohibit community college transfers into CS majors are likely to exacerbate existing disparities in CS enrollment.

### Participation in CS Education

We operationalize participation as enrolling in CS courses when offered by the school, either at the high school or college level. Examples of questions that address participation in CS education include:

- ▶ High School

- ▶ Are there enrollment disparities in advanced CS courses based on gender, geography, socioeconomic status, or ethnicity?

- ▶ College

- ▶ Are there disparities in CS majors based on student gender or ethnicity?

- ▶ Are male non-CS majors more likely than female non-CS majors to enroll in CS courses?

Undergraduate STEM majors in fields such as biological sciences are dominated by females. Given the increasing expectation that competency in these fields requires experience with computational and analytical tools grounded in computer science, the need to provide access for non-CS STEM majors in particular has implications for gender equity that should be examined.

### Experience of CS Education

Experience of CS education encompasses the various outcomes of participating in CS. The overarching

## Policies at four-year institutions that effectively prohibit community college transfers into CS majors are likely to exacerbate existing disparities in CS enrollment.

questions here are: When students participate in CS, do they have equitable learning experiences? What have they learned? Are their experiences culturally and personally relevant? Are students successful academically? Do all students feel welcome in the class?

Additional questions to assess student experiences of CS education include:

- ▶ High School

- ▶ Do course curricula explicitly address issues of equity?

- ▶ What is the relationship between AP test outcomes and gender and ethnicity?

- ▶ College

- ▶ Do passing rates or grades differ between student subpopulations based on demographics that should not be correlated with academic achievement?

- ▶ Do all students feel included and accepted in CS courses? Are females and students of color more likely to drop out of CS majors?

Student performance measures such as course grades, degree attainment, and AP test outcomes are one way to measure equitable outcomes for students, but providing truly equitable experiences must go beyond these simple outcome measures as it is possible to have parity in these types of outcomes while still failing to create an environment where all students feel they belong, instruction is inclusive, and diverse perspectives are valued explicitly. To achieve this inclusivity, instructors must attend to the explicit and implicit policies, classroom culture, and instructional strategies that either support or discourage underrepresented students in CS courses.

We argue that efforts to diversify the computing profession must use an ecosystems approach to account for the myriad contextual factors, institutional policies, and unexamined practices that influence the entire CS education pipeline. The CAPE framework can be a useful tool for examining some of the root causes that lead to a lack of diversity in computing. “If you build it, they will come” may work well for baseball movies, but diversifying computing education and the computing profession will require a more comprehensive examination of all levels of the CS ecosystem and the ways in which issues of equity, diversity, and inclusion play out to either exacerbate or mitigate existing disparities. The CAPE framework can be a road map for examining both the leading indicators of equity in CS, such as capacity and access, and the lagging indicators of student outcomes. Individuals committed to broadening participation in the computing field must be prepared to address each of these interrelated levels if we hope to build a more diverse computing profession. **C**

### References

1. Code.org, CSTA, and ECEP Alliance. *2020 State of Computer Science Education: Illuminating Disparities*. (2020); <https://bit.ly/3m6hHMh>
2. Computing Research Association (CRA). *Generation CS: Computer Science Undergraduate Enrollments Surge Since 2006*. (2017).
3. Ericson, B. *AP CS A and CSP Data*. Computing for Everyone. (Jan. 13, 2020); <https://bit.ly/3a3lGqA>
4. Ginder, S.A., Kelly-Reid, J.E., and Mann, F.B. *Enrollment and Employees in Postsecondary Institutions, Fall 2017, and Financial Statistics and Academic Libraries, Fiscal Year 2017: First Look (Provisional Data)* (NCES 2019-021rev). U.S. Department of Education. (2018); <https://bit.ly/2W8D7xK>
5. Google Inc. and Gallup Inc. *Diversity Gaps in Computer Science: Exploring the Underrepresentation of Girls, Blacks and Hispanics*. (2016); <http://goo.gl/PG34aH>
6. Hill, C., Corbett, C., and St. Rose, A. *Why So Few?: Women in Science, Technology, Engineering, and Mathematics*. American Association of University Women, (2010); <https://bit.ly/2KhNrRD>
7. Ma, J. and Baum, S. *Trends in Community Colleges: Enrollment, Prices, Student Debt, and Completion*. College Board, 2016; <https://bit.ly/3oJ8tr3>
8. Nelson, B. The data on diversity. *Commun. ACM* 57, 11 (Nov. 2014), 86–95; <https://doi.org/10.1145/2597886>

**Carol L. Fletcher** (cfletcher@tacc.utexas.edu) is the director of Expanding Pathways in Computing (EPIC) at The University of Texas at Austin's Texas Advanced Computing Center (TACC), Austin, TX, USA.

**Jayce R. Warner** (jwarner@tacc.utexas.edu) is a research associate for EPIC at The University of Texas at Austin's Texas Advanced Computing Center (TACC), Austin, TX, USA.

The development of the CAPE Framework was supported in part by a Google CS-ER Grant to The University of Texas at Austin.

Copyright held by authors.





George V. Neville-Neil

DOI:10.1145/3442375

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

## Kode Vicious Kabin Fever

*KV's guidelines for KFH (koding from home).*

**Dear KV,**

Forgive me if this seems off topic, but I was wondering if you had any advice for the majority of us who are now KFH (koding from home). I don't know how KV works day to day, but it seems pretty clear that the status quo has changed at most workplaces in the last several months, and it is difficult to know if there are things we could be doing to stay productive while we are all at home, ordering delivery, and microwaving our mail. Does KV have any good guidance?

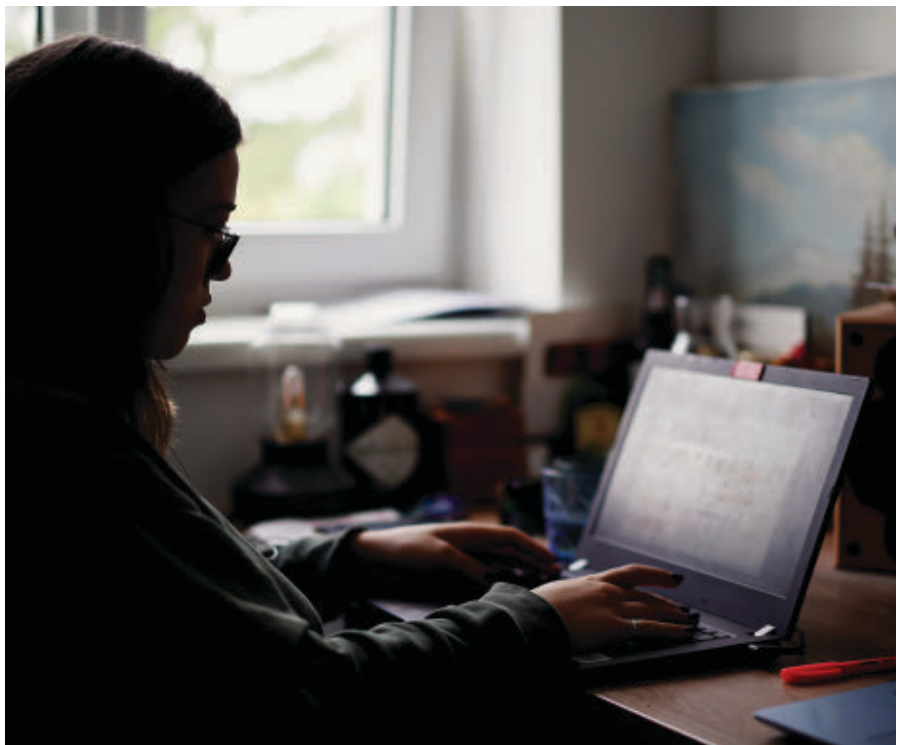
**Kabin Fever**

**Dear Kabin,**

Let me invite you to my next Zoom meeting on how to host Zoom meetings! Yes, like the rest of the world, KV has been koding from home—when he is not screaming from home or breaking furniture from home.

As a devotee of mobile computing and remote work from my earliest university days—where, for one of my co-op jobs, I worked on packaged software for the Commodore Amiga from my dorm room—I have, over time, developed a number of useful habits for maintaining a good and productive working rhythm, and I have found many of these apply well to those of you who are newly working from home. (One note: I do not now, nor have I ever, had children, so I will not address the complexities of working from home while you have kids in the house.)

Here are KV's guidelines for working from home.



► **Set your alarm and wake up at the same time each day.** I do not mean set the alarm for some ungodly hour, such as 8 A.M., unless that is when you would normally wake. I mean you want to keep a regular working schedule. During the university gig I mentioned, I worked from 8 P.M. until 4:00 A.M., five days a week, and then I slept until noon each day. That happens to be how I like to work, and that job did not demand any day-to-day interaction with co-workers; I only had to produce new versions of the software each Friday for review. If you work with a group of

other people, you should ensure you have some overlapping hours (two to three) with the majority of them, so meetings are possible.

► **Shower and dress as if you are going to the office as you normally would.** Many people think those of us who normally work from home do so in our pajamas. KV does not wear pajamas, ever, but he does put on pants and some sort of shirt every day. Do not underestimate the effect that a change of clothes will have on a change of your attitude toward work. If you work in your sleeping clothes, you are very likely going to have a

problem delineating work time from nonwork time.

► **Set a finishing time for each day and stick to it.** Keeping a proper life/work balance for someone who is used to going to an office is more difficult when you switch to working from home. Suddenly you do not have a commute and can roll from bed to desk and back.

► **Take frequent breaks of at least 15 minutes per two-hour block throughout the day.** We all love to be in the zone, but our eyes do not, and staring at a screen without interruption for eight to 10 hours a day is even easier at home where there are no coworkers to interrupt you.

► **Silence all your messaging apps.** Slack, IRC, Hangouts, and every other messaging app ever invented now cause a lot more interruptions than they did when you were in the office because everyone is now alone and cannot survive without the hallway conversations that lubricated their days. These apps are a major source of distraction and should be silenced, while leaving their counting badges on. When you take one of your 15-minute breaks, you can check these apps to see if anything of true importance lies there. The nice thing about ignoring them for long stretches at a time is that people will often have found the answers they are looking for on their own by the time you check, which saves you time and gives them a learning experience.

► **Do not use social media during your breaks.** “Doomscrolling” is problematic, and it is not the right way to take a break. Breaks are meant for getting up, walking around your cell (I mean home office), getting another coffee, maybe looking out the window ...

► **Arrange social time with actual humans outside of work.** Yes, we are all masked, wrapped in plastic, and supposed to wave at each other from the sealed confines of our homes, but one of the things that all humans need is human connection. Many companies have been hosting videoconferences, games, and other such activities during or after work hours, but I find these to be tedious and pointless as they are just like being trapped in yet another meeting with your coworkers.

**As a devotee of mobile computing and remote work from my earliest university days I have, over time, developed a number of useful habits for maintaining a good and productive working rhythm, and I have found many of these apply well to those of you who are newly working from home.**

A reasonable antidote to these distractions, if you cannot get out to a park to meet friends at a social distance, is to call a friend. Back in the old days we had these telephone things, and we would call friends and talk with them, sometimes about nothing at all. The sound of a friend’s voice on the phone is far more likely to keep you sane than a contrived game with your coworkers, with whom you might have just spent the day online.

► **Exercise.** I can hear you screaming for my blood now, but KV is an ardent cyclist and has found he is riding even more now that it is something he can do that is physically distanced from others and gets the blood moving. Gyms still seem to be problematic, but a walk in a park is not, so try that. When people went to offices at least they walked to and from the parking lot or from public transit. Now, you can literally take 100 steps per day and that, actually, is not good for you as it is not enough.

► **Use your old commute time to learn something new.** The average

person commutes two hours per day. You could give that time to your employer—who will happily take it from you—but you could also use that time to learn new skills, hack on a personal project, or read about a topic you want to know more about. Someday this pandemic will end, you will go back to commuting, and you will again lose those two hours, trading them for food and shelter, so take advantage of them now, while you have them.

► **If you schedule meetings, make them count.** One thing the pandemic has taught us is that most people do *not* know how to run a meeting, which, honestly, is going to be the topic of another KV column because I am nearly out of space here. In these troubled times, where everyone seems to want more status meetings, it is important to note meetings should be short, have an agenda, and be run with ruthless efficiency. That does not mean you cut people off without provocation, but it does mean you do not let meetings meander into unproductivity. Meetings are meant to share information quickly and, often, to arrive at a group consensus on solving a problem. Don’t be too shy to shut down meetings that are pointless. KV does this all the time, pandemic or not. Let’s face it, over the past six months we have all had enough of Zoom, Hangouts, and the like to last us several lifetimes.

I hope these tips and tricks help you now and serve you well if you continue to work from home after the current emergency is past. Best of health to all of you.

**KV**

**Q** Related articles on [queue.acm.org](https://queue.acm.org)

**The Paradox of Autonomy and Recognition**

Kate Matsudaira

<https://queue.acm.org/detail.cfm?id=2893471>

**CTO Roundtable: Cloud Computing**

<https://queue.acm.org/detail.cfm?id=1551646>

**A Conversation with Ray Ozzie**

<https://queue.acm.org/detail.cfm?id=1105674>

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

Copyright held by author.

## Viewpoint

# Cybersecurity: Is It Worse than We Think?

*Evaluating actual implementations and practices versus stated goals.*

**C**YBERSECURITY CONSISTENTLY RECEIVES significant attention, pressuring organizations to take precautionary steps to prevent incidents and data breaches. Numerous surveys are published each year by reputable organizations such as Deloitte, Verizon, The Ponemon Institute, and ISACA to get a better sense of what organizations are doing in response to these pressures. The general attitude is that threats evolve quickly and many organizations struggle to keep up.<sup>5</sup> Much of the data available on this subject comes directly from cybersecurity professionals, which provides legitimacy to the findings. However, it also represents a somewhat biased sample in that responding organizations have already committed resources to tackling these complex issues. Further, there is limited analysis on how individual organizations are changing over time as such reports typically provide industry-level observations. We seek to complement the myriad security research notes by investigating specific cybersecurity practices within organizations to evaluate where organizations are showing improvement, where they are stagnant, and what may be influencing these changes. Our results confirm that cybersecurity continues to receive attention on the surface, but when looking beyond surface-level impressions a surprising lack of progress is being made.

### Peeling Back the Layers

Each year, the Society for Information



Management (SIM) conducts the IT Trends Study—an extensive survey of CIOs and top IT executives to evaluate IT practices within organizations.<sup>1</sup> Organizations come from 30 different industries and vary in size, with an average revenue of \$4 billion and a median revenue of \$400 million. A hallmark of the study is the annual ranking of “organizations’ Top IT management Issues” where respondents are asked to select up to five IT-related issues from a list of 41 that are the “greatest concerns to their organization.” Cybersecurity has been in the top 10

for a decade as was the top concern for the last three years, signaling that organizations are more worried about cybersecurity than any other IT concern. However, the percentage of organizations selecting cybersecurity was only 41.9% in 2017, 38.3% in 2018, and 35.9% in 2019, suggesting a reality where a relatively small percentage of organizations treat it as a top concern.

One possible explanation of this decline is that significant cybersecurity improvements have already been made, shifting organizational priorities elsewhere. To better evaluate



whether this is the case, we ask respondents whether their organization:

- ▶ Has a CISO or equivalent?
- ▶ Requires cybersecurity training for employees?
- ▶ Considers cybersecurity during software development, change management, IT procurement, and/or overall business strategy?
- ▶ Measures and evaluates cybersecurity performance?
- ▶ Has cyber insurance coverage?

While these questions do not provide absolute assurance that an organization is adequately prepared to address all cybersecurity threats, they do provide the opportunity to see how organizations are changing over time (since many respondents participate in multiple years). Additionally, negative responses signal that an organization is clearly not adopting common cybersecurity best practices. In comparing 2016 to 2019, it is clear there is improvement in some areas yet growth is stagnant in others (see the table here).

The most dramatic change comes in the form of cyber-insurance. Fewer than half of organizations had such coverage in 2016 but nearly two-thirds were covered in 2019. By transferring risk to a third party, an organization may focus on other top priorities. However, cyber-insurance is by no means a panacea as it will typically not provide financial compensation for lost sales, reputational damage, or costs associated with fortifying systems.<sup>2</sup> For example, Target estimated the financial impact of their breach in 2013 was \$291 million but only \$90 million was offset through insurance coverage.<sup>6</sup> While the significant increase in companies adopting cyber-insurance plans is admirable, in the absence of other significant security improvements, it may provide limited risk reduction for organizations.

Cybersecurity's involvement in the IT Procurement process has also seen a notable improvement since 2016. Given the rise in cloud utilization<sup>1</sup> and the interconnectedness of vendor/supplier systems, risk exposure continues to expand outward from the organization. Further, 59% of breaches in 2018 involved third-party systems or failures.<sup>3</sup> As such, it appears as though organizations are placing more emphasis on en-

## Cyber-insurance is by no means a panacea as it will typically not provide financial compensation for lost sales, reputational damage, or costs associated with fortifying systems.

suring adequate security provisions are included when purchasing IT components or engaging third parties.

Despite the improvements noted over the past four years, there is still room for growth. The figures noted in the last row in the table represent a sum of overall readiness that is determined by awarding one point for an affirmative answer to each of the five questions included in the survey (organizations received 0.25 points for each business process security was integrated with). While gradual improvement has been observed over the last four years, the average organization still only implemented 3 out of 5 standard best practices in place in 2019.

While our sample is skewed toward small to medium-sized organizations,

large companies are also experiencing issues. Of organizations with revenues greater than \$1 billion in 2019 (30% of our sample), only 75.2% had a CISO or equivalent position and the average readiness score was a 3.51. For organizations over \$5 billion in revenue (13% of our sample), 81.4% had a CISO and the average readiness score was 3.57. While large companies are in a slightly better position, there is still room for improvement.

So, is cybersecurity worse than we think? We think the answer is yes—after peeling back the layers to identify specific practices within organizations, there is much to be desired. With approximately 50% of organizations appointing a leader of cybersecurity efforts and involving security in the planning of overall business strategy, many organizations, even the ones with respectable readiness scores, are tackling cybersecurity as more of an IT process rather than an enterprisewide issue. Of course, simply appointing a leader or providing a seat at the table for strategy planning meetings is not effective unless the organization truly buys into the importance of cybersecurity.

### Prioritizing Cybersecurity

To better understand the impact of setting the tone at the top with a focus on cybersecurity, we were curious as to whether organizational prioritization has any effect on cybersecurity practices. We compared cybersecurity readiness scores and organizational prioritization across a two-year period for those organizations that provided responses in two consecutive years.

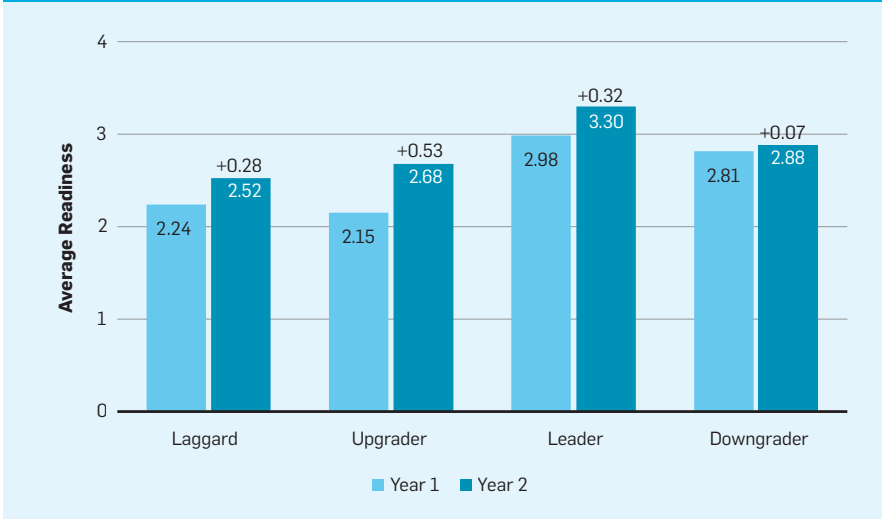
#### Cybersecurity Practices: 2016 vs. 2019.

Cybersecurity Practice	2016 (n=685)	2019 (n=501)	Relative % Change in three years*
CISO	45.8%	48.8%	6.6%
Cybersecurity Involvement In:			
IT Procurement	53.6%	68.1%	27.1%**
Software Development	79.3%	78.2%	-1.4%
IT Change Management	79.0%	73.9%	-6.5%**
Overall Business Strategy	49.1%	52.3%	6.5%
Mandatory Security Training	61.5%	77.6%	26.2%**
Cyber-Insurance	47.6%	65.9%	38.4%**
Cybersecurity Performance Measures	21.4%	29.5%	37.7%**
Overall Readiness	2.54	2.94	15.8%

\* Relative percentage change is calculated by dividing the raw percentage differences between the two years by the percentage in 2016. For example,  $(48.8\% - 45.8\%) / 45.8\% = 0.066$  (6.6%).

\*\* Statistically significant difference ( $p < 0.05$ , chi-squared test of proportions)

Change in readiness across classes.



Each organization was classified into one of four classes:<sup>a</sup>

- ▶ “Leaders”: an organizational priority in both years (28.0% of organizations)
- ▶ “Laggards”: not an organizational priority in either year (37.9%)
- ▶ “Upgraders”: an organizational priority in year two but not in year one (17.6%)
- ▶ “Downgraders”: an organizational priority in year one but not in year two (16.4%)

The readiness scores (see the figure here) reveal two statistically significant insights.<sup>b</sup> First, organizations that prioritize cybersecurity have higher readiness scores. Leaders rise above the rest, with downgraders close behind. Second, improvements to cybersecurity readiness are different across these classes. When comparing the classes, we see the worst-performing class, in terms of improvement, is the “downgrader” (+0.07) whereas “upgraders” resulted in the largest one-year improvement (+0.53). This suggests that organizations that turn their attention away from cybersecurity see virtually no improvement whereas those that make a conscious decision to begin

treating it as a priority observe much greater improvements.

These results offer only a two-year snapshot and it is common knowledge that improvements to cybersecurity defenses take time. For the 139 organizations we have 36 months of data for, the improvements from year one to year two were almost identical to the year two to year three improvements across all four classes. Thus, the pace by which improvements are observed is steady across multiple years.

### What Does It All Mean?

Given our analysis, we believe there is a harsh reality lurking beneath the surface within many organizations. While they may be saying the right things in public to satisfy investors, underwriters, and customers, there is an apparent lack of urgency in promoting a truly resilient and secure organization. Our research did not have to dig very deep to find surprising gaps in organizational security practices. Further, the security practices most commonly missing from organizations tend to be those that provide visibility, leadership, and integration with the business.

Our data also suggests when organizations say cybersecurity is one of their top concerns, they tend to do more about it. However, they still appear to be reluctant to hire a CISO or provide cybersecurity a seat at business strategy meetings. Our data suggests large companies are doing better in this regard, but even they still

struggle to implement all of these foundational security practices.

Why is this the case? Although we cannot objectively answer this question, we can offer several possible conjectures. First, cybersecurity budgets are notoriously difficult to justify given there is no true ROI.<sup>4</sup> Hiring a CISO is a large investment whereas developing a short training video or document and distributing it to all employees requires minimal financial resources. Second, it is possible that risk tolerances of CEOs may be rising. Given changes to compensation structures for top executives and pressures from investors to deliver short-term gains, there is little incentive to divert resources away from ventures that deliver near-term returns. As such, CEOs may be wary of inviting security personnel into strategic planning discussions for fear of security requirements inhibiting productivity and innovation. Finally, it is possible that a defeatist mentality is setting in across organizations. Everyone has heard the phrase “it’s not a matter of if, but when” in terms of cybersecurity incidents so perhaps organizations are simply doing the bare minimum and are prepared to face the consequences when the inevitable occurs.

Cybersecurity threats are not going anywhere and even well-prepared organizations will continue to experience breaches. This does not mean we should give up, however. Organizations of all shapes and sizes have plenty of room for improvement once you look beneath the surface. ■

### References

1. Kappelman, L. et al. The 2018 SIM IT issues and trends study. *MIS Quarterly Executive*. (2019).
2. Mak, A. *AdvisorSmith* 18, 7 (2019); <https://bit.ly/3a2IIDq>
3. Ponemon Institute. (Nov. 2018); <https://bit.ly/2W4f8zI>
4. Schneier, B. *CSO* 9, 2 (Nov. 2008); <https://bit.ly/3qMpMta>
5. Staff, C. *Cybersecurity*. *Commun. ACM* 60, 4 (Apr. 2017).
6. Target. *SEC Edgar* 5, 25 (2016); <https://bit.ly/37SIN4t>

**Chris Maurer** (maurer@virginia.edu) is an assistant professor in the McIntire School of Commerce at the University of Virginia in Charlottesville, VA, USA.

**Kevin Kim** (kevin.kim@unt.edu) is a Ph.D. student in the G. Brint Ryan College of Business at the University of North Texas in Denton, TX, USA.

**Dan Kim** (Dan.Kim@unt.edu) is a professor in the G. Brint Ryan College of Business at the University of North Texas in Denton, TX, USA.

**Leon Kappelman** (kapp@unt.edu) is a professor in the G. Brint Ryan College of Business at the University of North Texas in Denton, TX, USA.

Copyright held by authors.

<sup>a</sup> 414 unique organizations participated in the survey in at least two consecutive years between 2016 and 2019, 139 organizations provided at least three consecutive years of data, and 43 organizations provided four years of data.

<sup>b</sup> A mixed model controlling for organization and sector was evaluated. The difference in readiness based only on organizational concern (0 or 1) was statistically significant at the 0.01 level. Differences in readiness between prioritization classes was significant at the 0.05 level.

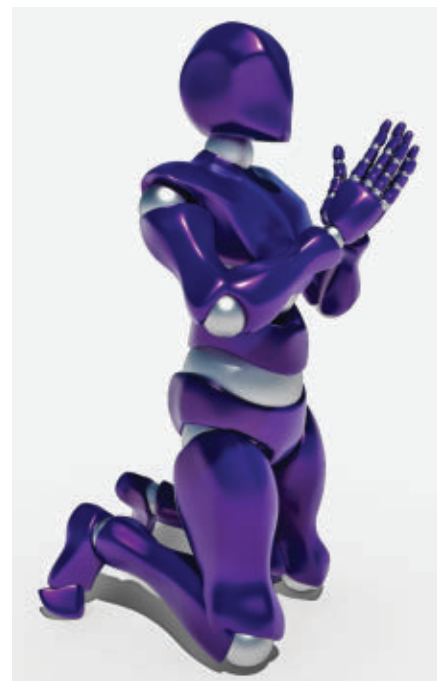
## Viewpoint

# Polanyi's Revenge and AI's New Romance with Tacit Knowledge

*Artificial intelligence systems need the wisdom to know when to take advice from us and when to learn from data.*

**I**N HIS 2019 Turing Award Lecture, Geoff Hinton talks about two approaches to make computers intelligent. One he dubs—tongue firmly in cheek—“Intelligent Design” (or giving task-specific knowledge to the computers) and the other, his favored one, “Learning” where we only provide examples to the computers and let them learn. Hinton’s not-so-subtle message is that the “deep learning revolution” shows the only true way is the second.

Hinton is of course reinforcing the AI zeitgeist, if only in a doctrinal form. Artificial intelligence technology has captured popular imagination of late, thanks in large part to the impressive feats in perceptual intelligence—including learning to recognize images, voice, and rudimentary language—and bringing fruits of those advances to everyone via their smartphones and personal digital accessories. Most of these advances did indeed come from “learning” approaches, but it is important to understand the advances have come in spheres of knowledge that are “tacit”—although we can recognize faces and objects, we have no way of articulating this knowledge explicitly. The “intelligent design” approach fails for these tasks because we really do not have conscious theories for such tacit knowledge tasks. But, what of tasks and domains—especially those we de-



**“Human, grant me the serenity to accept the things I cannot learn, data to learn the things I can, and wisdom to know the difference.”**

signed—for which we do have explicit knowledge? Is it forbidden to give that knowledge to AI systems?

The polymath Polanyi bemoaned the paradoxical fact that human civilization focuses on acquiring and codifying “explicit” knowledge, even though a significant part of human knowledge is “tacit” and cannot be exchanged through explicit verbal instructions.

His “we can know more than we can tell” dictum has often been seen as a pithy summary of the main stumbling block for early AI efforts especially in perception.

Polanyi’s paradox explains to a certain extent why AI systems wound up developing in a direction that is almost the reverse of the way human babies do. Babies demonstrate aspects of perceptual intelligence (recognizing faces, voices and words), physical manipulation (of putting everything into their mouths), emotional intelligence, and social intelligence, long before they show signs of expertise in cognitive tasks requiring reasoning skills. In contrast, AI systems have demonstrated reasoning abilities—be they expert systems or chess—long before they were able to show any competence in the other tacit facets of intelligence including perception.

In a sense, AI went from getting computers to do tasks for which we (humans) have explicit knowledge, to getting computers to learn to do tasks for which we only have tacit knowledge. The recent revolution in perceptual intelligence happened only after labeled data (such as cats, faces, voices, text corpora, and so forth) became plentiful, thanks to the Internet and the World Wide Web, allowing machines to look for patterns when humans are not quite able to give them explicit know-how.



Lately though, Polanyi's paradox is turning into Polanyi's revenge both in research and practice of AI. Recent advances have made AI synonymous with learning from massive amounts of data, even in tasks for which we do have explicit theories and hard-won causal knowledge.<sup>a</sup> This resistance to accept any kind of explicit knowledge into AI systems—even those operating in tasks and environments of our design—is perplexing. The only “kosher” ways of taking explicit knowledge in deep learning systems seem to be to smuggle them in through architectural biases, or feeding them manufactured examples. Anecdotal evidence hints that industry practitioners readily convert doctrine and standard operating procedures into “data” only to have the knowledge be “learned back” from that data. Even researchers are not immune—a recent paper in *Nature Machine Intelligence* focused on how to solve Rubik's Cube by learning from billions of examples, rather than accept the simple rules governing the puzzle. There are policy implications too. Many governmental proposals for AI research infrastructure rely exclusively on creating (and curating) massive datasets for various tasks.

The current zeal to spurn hard-won explicit (and often causal) knowledge, only to try to (re)learn it from examples and traces as tacit knowledge, is quixotic at best. Imagine joining a company, and refusing to take advice on their standard operating procedures, and insisting instead on learning it from observation and action. Even if such an approach might unearth hidden patterns in how the company actually works, it will still be a wildly inefficient way to be an employee. Similar concerns will hold for AI assistants in decision support scenarios.

A common defense of this “learning first” trend is the asymptotic argument that since we humans—with an

essentially neural basis of their brains—have managed to develop shared representations and ability to communicate via explicit knowledge, AI systems based purely on learning may well be able to get there eventually. Perhaps. But it is quite clear that we are far from that point, and a misguided zeal to steer away from AI systems that accept and work with explicit knowledge is causing a plethora of problems right now.

Indeed, AI's romance with tacit knowledge has obvious adverse implications to safety, correctness, and bias of our systems. We may have evolved with tacit knowledge, but our civilization has been all about explicit knowledge and codification—however approximate or aspirational. Many of the pressing problems being faced in the deployment of AI technology, including the interpretability concerns, the dataset bias concerns as well as the robustness concerns can be traced rather directly back to the singular focus on learning tacit knowledge from data, unsullied by any explicit knowledge taken from the humans. When our systems learn their own representations from raw data, there is little reason to believe that their reasoning will be interpretable to us in any meaningful way. AI systems that refuse to be “advised” explicitly are taking the “all rules have exceptions” dictum to the “what are rules?” extreme, which flies in the face of civilizational progress, and seriously hinders explainability and contestability of machine decisions to humans in the loop.

How confident can we be of a medical diagnostic system using AI, when it shares little common knowledge beyond raw data with the presiding physician? This is no longer a hypothetical. Just recently, a paper in *JAMA Dermatology* showed that a commercially approved AI system for melanoma detection was easily misled by surgical skin markings next to benign moles. Wittgenstein was alluding to this at some level, when he remarked “if a lion could speak, we could not understand him.”

At least part of the problem, in terms of public perceptions, is our own very human romance with tacit knowledge, which continues despite the fact that the progress of civilization depended on explicit knowledge. We tend to romanticize *je ne sais quoi* and ineffability (no one ever impressed their life mate by “explaining” their love with a crisp

numbered list of falsifiable attributes!). This very human trait makes feats of AI systems that learn without being told all that much more fascinating to us (nevermind their inscrutability and attendant headaches).

While we are easily impressed at computer performance in tasks where we have no conscious models and explicit knowledge (for example, vision, speech), there are also many domains, especially those designed by humans, where we do have models and are willing to share them! Indeed, the hallmark of human civilization has been a steady accumulation of such explicit knowledge. After all, many animals have perceptual abilities that are more acute than we humans have, but we got much farther because of our ability to acquire and use explicit knowledge, rather than learn only from observation. It is important for AI systems to be able to take such knowledge when it is readily available, rather than insist on rediscovering it indirectly from examples and observation. There should be no shame in widening the pipeline between humans and AI systems and accepting readily offered knowledge, be it explicit norms and rules, causal models or shared vocabulary.

Of course, combining learning and explicit knowledge in a fully principled way continues to be an open problem. Often the explicit knowledge may only provide an initial bias that gets refined through learning. To do this effectively, we will need to go beyond ways to smuggle knowledge through model architectures. While we are busy trying to make headway on that problem however, we should at least resist the temptation to stigmatize acquisition and use of explicit knowledge.

We found it to be fruitless to insist on explicit knowledge for tacit tasks such as face recognition. It will be equally futile to ignore readily available explicit knowledge and insist on learning/recovering it from examples. Our machines must have the wisdom to know when to take advice and when to learn. □

---

**Subbarao Kambhampati** (rao@asu.edu) is a professor of computer science at Arizona State University, Tempe, AZ, USA. He is the past president of the Association for the Advancement of Artificial Intelligence, and a fellow of AAAI, AAAS, and ACM. He can be followed on Twitter at @rao2z. A longer talk on this topic is available at <https://bit.ly/3kyUNND>.

Copyright held by author.

<sup>a</sup> The recent interest in taking deep learning systems beyond their current reflexive System 1 capabilities to deliberative System 2 ones is related, but somewhat orthogonal to the tacit/explicit knowledge distinction. While most tacit knowledge tasks do get handled at System 1, explicit knowledge tasks start in System 2 but may get compiled into System 1 reflexive behavior for efficiency. My interest here is in having AI systems leverage human know-how on explicit knowledge tasks.

► Len Shustek, Column Editor

## Viewpoint

# Let's Not Dumb Down the History of Computer Science

*Donald Knuth on the best way to recognize the history of computer science.*

**Editor's note:** On May 7, 2014, Don Knuth delivered that year's Kailath Lecture<sup>a</sup> at Stanford University to a packed auditorium. In it he decried the absence of technical content from the histories of computer science being written, and he made an impassioned plea for historians of computer science to get back on track, as the historians of mathematics have always been.

Both the video<sup>b</sup> and, now, the verbatim transcript<sup>c</sup> of that talk are online. In the January 2015 issue of *Communications*,<sup>d</sup> historian Thomas Haigh analyzed and responded to the talk, concluding that “work of the particular kind preferred by Knuth will flourish only if his colleagues in computer science are willing to produce, reward, or commission it.”

This Viewpoint, which we thank *Communications* Senior Editor Moshe Vardi for suggesting, is a condensed and highly edited transcript of the original talk that has provoked so much discussion.

a See <https://stanford.io/3qYDCce>

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

c See <https://stanford.io/2Wg2v4J>

d “The Tears of Donald Knuth,” Thomas Haigh, *Commun. ACM* 58, 1 (Jan. 2015), 40–44; <https://bit.ly/382aAQ7>



**G**IVING THIS TALK might be the greatest mistake in my life, because I'm going to talk about controversial things. I generally go out of my way to avoid argument whenever possible. But I feel so strongly about this that I just have to vent and say it.

Although there is “history” in the title, I'm not going to tell you about the history of computer science. Instead,

I'm going to talk about historians of computer science—about historiography. This is meta-history. I'm going to try to explain why I love to read works on history, and why I'm profoundly disturbed by recent trends in what I've been reading.

Why do I, as a scientist, get so much out of reading the history of science? Let me count the ways:

1. To understand the process of

discovery—not so much what was discovered, but how it was discovered. Primary sources are best: the words of somebody who discovered something, as they were discovering it. The more examples I see, the more likely I'll be able to discover something tomorrow.

2. To understand the process of failure. We learn a good deal from historical errors, not only from our own. It also helps to know that even the greatest minds are unable to grasp things that seem obvious to us. Leibniz spent much time working on combinatorics, and most of what he did was underwhelming and totally wrong.

3. To celebrate the contributions of many cultures. There are many ways of thinking, many points of view, and many independent researchers. Fibonacci numbers were discovered in India long before Fibonacci. Catalan numbers were discovered in China, a hundred years before Catalan. Many uneducated people have discovered wonderful patterns in numbers, and I can share their joy of discovery.

4. Telling historical stories is the best way to teach. It's much easier to understand something if you know the threads it is connected to. Give credit to Fibonacci, but also to Narayana in India. The complete story is of many separate individuals building a magnificent edifice with a series of small steps.

5. To learn how to cope with life. How did other scientists grow up, make friends or enemies, manage their time, find mentors, mentor others, and serve their communities? Balance is important.

6. To become more familiar with the world, and to know how science fits into the overall history of mankind. What was life like on different continents and in different epochs? The main difference between human beings and animals is that people learn from history.

I am grateful in particular to historians of mathematics. They make original source materials accessible through reprints, and through their translations of both language and notation. They scout out previously unpublished papers, letters, meeting minutes, and official records, and then link them together into a narrative. What I don't like is analysis of trends alone; I like to see the source materials up front.

So there is mostly good news from the historians of mathematics. The bad news comes from the historians of computer science.

What did it for me was an article by Martin Campbell-Kelly, a leading historian of computer science whose work I had admired. But his 2007 article on “The History of the History of Software”<sup>e</sup> was a shock.

The centerpiece of the article was a table that classified selected works on software from 1967 to 2004 into four categories: technology; industry; applications; and institutional/social/political. At the beginning most published works are about the technology, but by the end they are mostly in the other categories. The author's description of the change is that “over time, software history has evolved from narrow technical studies, through supply-side and economic studies, to broad studies of applications.”

He thinks that is good! On the contrary, it is extremely shallow and completely non-technical. I broke down and started to cry. I finished reading it only with great difficulty because tears had made my glasses wet. I immediately dashed off a letter to Martin.<sup>f</sup>

“I must confess that by the time I got to the last three or four pages, I was so upset that I could barely see straight. I had to force myself to read slowly, not believing you had succumbed so far to the alarming-to-me trends and fads of the moment about how history ‘ought to be’ written.

“Do you not see any blind spots in your outlook when your Table 1 shows 68% class T [technology] articles in the first 20 years, and 0% class T in the last five years ... and then you say ‘The table shows how the subject matter has broadened!’ The subject matter has not broadened; it has totally shifted. All we get nowadays is dumbed-down. Thank goodness historians of mathematics have not entirely abandoned writing articles that contain formulas or explain scientific ideas.

<sup>e</sup> “The History of the History of Software,” Martin Campbell-Kelly, *IEEE Annals of the History of Computing* 29, 4 (Oct.–Dec. 2007); 40–51; <https://bit.ly/3oMC0jN>

<sup>f</sup> Campbell-Kelly replied in “Knuth and the Spectrum of History,” *IEEE Annals of the History of Computing* 36, 3 (July–Sept. 2014); <https://bit.ly/3ninEXP>

“I am sure that business histories are as difficult to write as technical histories, and they are no doubt also as valuable to businessmen as technical histories are valuable to technicians. But you seem to be celebrating the fact that nobody writes technical CS history at all anymore!

“When you speak of ‘obvious holes’, you are thinking of obvious holes in business history ... the video game industry, for example. But how about the people who write video games: They invent marvelous breakthroughs in techniques about how to render scenes and pack data and do things in parallel and coordinate thousands of online users. The lack of anything even close to describing these techniques and how they were discovered and under what constraints seems to me a much more obvious hole; yet you show no inclination to admit its existence much less to suggest plugging it.”

Martin wasn't always that way. He describes in the article how, for his Ph.D. dissertation under Brian Randell at Newcastle University, he “managed to locate most of the system programs developed for the first three operational British computers—the Cambridge EDSAC, the Manchester Mark I, the National Physical Lab Pilot ACE. Studying these programs and their texts was utterly absorbing.” Absolutely! He could see why it was beautiful. He was doing the kind of history that I came to admire him for.

Then by 1976 he was starting to think about the broader picture. He didn't see how it was “concrete” the way subroutine linkage was achieved on the EDSAC, or how you got an index register in the hardware of a machine. He offers a “biographical mea culpa” and says, “what they (we) wrote looks constrained, excessively technical, and lacking in breadth of vision.” He's apologizing for what I always had admired!

Back to my letter:

“During the past 20 years, histories and expositions of mathematics for general readers have gotten dramatically better, while the analogous histories and expositions of computer science have gone downhill. With your Table 1 you could have generated a wakeup call. But instead you seem to be a pied piper for continuing the dismal trends. You have clearly lost faith



in the notion that computer science is actually scientific (as well as being related to economics and defense etc.). Yet I still cling to that old-fashioned belief ... indeed, if computer science were no longer a rich science with deep ideas, I could finish *The Art of Computer Programming* in no time, but it appears that I still have 20 years of work ahead!”

Well, that was 5 years ago,<sup>g</sup> and I have 25 years of work ahead.

“You kindly state that it was OK and even fine for narrow-minded people like me to attempt to write history even though we have no training as historians, since there is a shortage of historians. Fair enough. But now you are encouraging professional historians to address only the masses of readers [...] and to ignore the 2% of the population who will spend their lives actually writing software. This you say is holistic and integrative. I view it as lightweight, mildly interesting; a chance to be witty and win some arguments so that another witty historian can challenge you and publish more lightweight stuff. Fine for employment of historians, but pretty much a waste of time for a reader who wants to know how to do hard science. The few papers I’ve written that have a historical component were among the most difficult I have ever done, and I greatly admire the historians who do it properly.”

I met Martin a few months later at a history meeting in England. We talked for several hours, but neither of us could get the other to agree. He keeps insisting that he wants his students to write no more books and papers of type T. Going back to my list of all the reasons why I love history, he’s saying that numbers one, two, three, and four aren’t important; only numbers five and six are of value.

I soon found out that historians of science have been debating this among themselves for a long time. They don’t call it “type T” versus something else; they talk about “internal history” versus “external history.” For them, internal history is written by and for people who are knowledgeable about some discipline, and the external histories are written for the masses. Internal histories, those of type T, have basically

**I am grateful in particular to historians of mathematics. They make original source materials accessible through reprints, and through their translations of both language and notation.**

come into disrepute—except, I’m glad to say, with respect to mathematics.

How has mathematics managed to escape this so far? I suppose it’s because historians of math have always faced the fact that they won’t be able to please everybody. Historians of other sciences have the delusion that any ordinary person can understand it, or at least they pretend so.

There was one thing that Martin Campbell-Kelly and I definitely agreed on: that it would really be desirable if there were hundreds of papers on history written by computer scientists about computer science. Specialists like me are not writing the kind of papers that would fill the historical gaps. Martin says at least he wants professional historians to have some data from misguided people—like we who do the technical stuff—that they can clean up later.

He muses about why it is that there are almost no history papers being written now by computer scientists, and he says that it is probably peer pressure—that papers on history don’t get any academic points. In Britain they had the notorious “Research Assessment Exercise,” which was used to decide on salaries and promotions. History papers probably got no points in that assessment, and so nobody writes them. In America I don’t see support for such

papers either. I think it’s something that computer scientists ought to do anyway, even though it’s hard to write these historical papers, and hard to get exposure for them.

I want to end on a high note, with a tantalizing wish list about what we could do. The best way to write history is to combine breadth and depth. Not just the broad ideas from which you understand the context, but also to zoom in on a few places and provide specific examples with detailed analyses. Here are some of the many papers waiting to be written:

► **Operating Systems.** I have at home Edsger Dijkstra’s source code for the operating system he wrote in 1965. Nobody has looked at it, and we should.

► **Databases.** Early computer programs were filled with database ideas that have never really been analyzed and placed in context.

► **Rendering techniques for movies and video games.** Many great technical ideas were developed at Pixar and elsewhere, and you could make a great story about the history of the algorithms they’ve used.

► **Compilers.** In the early 1960s there were really interesting programs written at Burroughs and Computer Sciences Corporation that have never been analyzed. There was a brilliant programmer at Digitek who had completely novel and now unknown ideas for software development; he never published anything, but you could read and analyze his source code.

► **The Computer History Museum has Bill Atkinson’s source code, now released by Apple, for MacPaint and MacDraw.** They are brilliant programs, beautifully organized and structured, that are a treat to read and deserve to be annotated and studied.

And so on. There are many wonderful algorithms and source codes whose histories are completely untouched. If we technicians can study and explain them in depth, then historians will at least have material to which they can later add the breadth. □

**Donald E. Knuth** is Professor Emeritus of The Art of Computer Programming at Stanford University, Stanford, CA, USA.

**Len Shustek** (len@shustek.com) is Chairman Emeritus at the Computer History Museum, Mountain View, CA, USA.

Copyright held by authors.

<sup>g</sup> 11 years now—Ed.

Article development led by **acmqueue**  
queue.acm.org

**A discussion with Miguel Guevara, Damien Desfontaines, Jim Waldo, and Terry Coatta**

# Differential Privacy: The Pursuit of Protections by Default

OVER THE PAST decade, calls for better measures to protect sensitive, personally identifiable information have blossomed into what politicians like to call a “hot-button issue.” Certainly, privacy violations have become rampant and people have grown keenly aware of just how vulnerable they are. When it comes to potential remedies, however, proposals have varied widely, leading to bitter, politically charged arguments. To date, what has chiefly come of that have been bureaucratic policies that satisfy almost no one—and infuriate many.

Now, into this muddled picture comes differential privacy. First formalized in 2006, it’s an approach based on a mathematically rigorous definition of privacy that allows formalization and proof of the guarantees

against re-identification offered by a system. While differential privacy has been accepted by theorists for some time, its implementation has turned out to be subtle and tricky, with practical applications only now starting to become available. To date, differential privacy has been adopted by the U.S. Census Bureau, along with a number of technology companies, but what this means and how these organizations have implemented their systems remains a mystery to many.

It’s also unlikely that the emergence of differential privacy signals an end to all the difficult decisions and trade-offs, but it does signify that there now are measures of privacy that can be quantified and reasoned about—and then used to apply suitable privacy protections.

A milestone in the effort to make this capability generally available came in September 2019 when Google released an open source version of the differential privacy library that the company has used with many of its core products.

In the exchange that follows, two of the people at Google who were central to the effort to release the library as open source—Damien Desfontaines, privacy software engineer; and Miguel Guevara, who leads Google’s differential privacy product development effort—reflect on the engineering challenges that lie ahead, as well as what remains to be done to achieve their ultimate goal of providing privacy protection by default. They are joined in this discussion by Jim Waldo, Harvard’s CTO who recently co-chaired a National Academies study on privacy, and Terry Coatta, the CTO of Marine Learning Systems.

**JIM WALDO:** I’d love to hear how you characterize differential privacy, since most of the descriptions I’ve heard so far are either so loose as to be meaningless or so formal as to be difficult to follow.

**MIGUEL GUEVARA:** I think about it in the context of other privacy technologies,







DAMIEN DESFONTAINES

**What's most characteristic about differential privacy is that when you generate statistics—that is, some aggregated information about a set of people—you purposely add noise to the results of that computation.**



many of which are policy- and heuristics-driven. That can make you feel good, but it's very hard to reason about a lot of those technologies, whereas differential privacy gives you a tangible way to reason about what's happening with the privacy of the underlying data and to quantify how much privacy has been lost there.

Having that ability is powerful for data curators. It also allows us to imagine a world where users possess that same sort of control over their own data and, by way of some adjustments to their applications, will have the ability to determine how much privacy they can have. So, the basic idea behind differential privacy is to give individuals the ability to make these sorts of decisions in a rational and informed manner.

**WALDO:** That's a nice way to characterize the goals of differential privacy. But now I'm going to ask you to get a little more concrete and talk about how you intend to meet those goals.

**DAMIEN DESFONTAINES:** What's most characteristic about differential privacy is that when you generate statistics—that is, some aggregated information about a set of people—you purposely add noise to the results of that computation. This is how you attain the guarantee of differential privacy: by ensuring that someone looking at the results of that computation will not be able to get information about the individuals whose data has been included as part of the dataset.

What I mean by *noise* simply has to do with sampling a random number of data points from a distribution. Ideally, that random number can be kept quite small—between -10 and 10 for a count, for example. For statistics on a larger scale, the noise you add should not greatly impact the quality of your data. Then, as Miguel indicated, differential privacy also lets you quantify the trade-offs between privacy and precision for a dataset. The amount of noise you add to the data is what allows you to quantify just how private the dataset will be. Which is to say, the more noise you add, the less precise your statistics will be. At the same time, your privacy guarantees will also become that much stronger.

**WALDO:** So, the core idea is that when you query the data, the answer has

some noise added to it, and this gives you control over privacy because the more noise you add to the data, the more private it becomes—with the trade-off being that the amount of precision goes down as the noise goes up.

**DESFONTAINES:** That's right.

**WALDO:** How is this now being used inside Google?

**GUEVARA:** It's mostly used by a lot of internal tools. From the start, we saw it as a way to build tooling that could be used to address some core internal use cases. The first of those was a project where we helped some colleagues who wanted to do some rapid experimentation with data. We discovered that, much of the time, a good way to speed access to the data underlying a system is to add a privacy layer powered by differential privacy. That prompted us to build a system that lets people query underlying data and obtain differentially private results.

After we started to see a lot of success there, we decided to scale that system—to the point where we're now building systems capable of dealing with data volumes at Google scale, while also finding ways to serve end users, as well as internal ones. For example, differential privacy made it possible for Google to produce the COVID-19 Community Mobility Reports [used by public health officials to obtain aggregated, anonymized insights from health-care data that can then be used to chart disease movement trends over time by geography as well as by locales (such as grocery stores, transit stations, and workplaces)]. There's also a business feature in Google Maps that shows you how busy a place is at any given point in time. Differential privacy makes that possible as well. Basically, differential privacy is used by infrastructure systems at Google to enable both internal analysis and some number of end-user features.

**WALDO:** As I understand it, there's a third variable. There's how accurate things are and how much noise you add—and then there's the number of queries you allow. Do you take all three of those into account?

**GUEVARA:** It really depends on the system. In theory, you can have an infinite number of queries. But there's a critical aspect of differential privacy called the privacy budget—each time you use

a query, you use some part of your budget. So, let's say that every time you issue a query, you use half of your remaining budget. As you continue to issue more queries, the amount of noise you introduce into your queries will just increase.

With one of our early systems, we overcame this by doing something you're hinting at, which was to limit the number of queries users could make. That was so we wouldn't exhaust the budget too fast and would still have what we needed to provide meaningful results for our users.

**DESFONTAINES:** There's also a question that comes up in the literature having to do with someone using an engine to run arbitrary queries over a dataset—typically whenever that person does not have access to the raw data. In such use cases, budget tracking becomes very important. Accordingly, we've developed systems with this in mind, using techniques like sampling, auditing, and limiting the number of queries that can be run. On the other hand, with many common applications, you know what kind of query you want to run on the data: For a busyness graph displayed on Google Maps, for example, a handful of predetermined queries might be used daily to generate the required data so you don't have to provide a higher privacy budget for future queries as yet unknown. Instead, you'll know in advance which queries are going to be issued, so you'll also know how much noise needs to be added.

**TERRY COATTA:** It seems a corollary of this might be: If you have a dataset against which you intend to perform ad hoc queries but don't know in advance what the nature of those queries might be, differential privacy in some sense limits the utility of that dataset. That is, there are only so many ad hoc queries that can be served before you've effectively exhausted your ability to query anymore against that dataset.

**GUEVARA:** OK, but I guess I would frame this in terms of use cases. What we've discovered is that when you look at the sorts of use cases you're suggesting, people tend to be interested in looking only at broad statistical trends. Say some company just introduced its product in Country X and now wants to see how many users are

using operating system 1 versus operating system 2. At that level, differential privacy provides really good results from a statistical perspective.

But then there's another use case, which is what I believe Damien was talking about. Let's say that, for this same example, you discover that the critical variables for your analysis happen to be country, age, and income. You can just set up a query accordingly and then run that every day or every few days without consuming any additional privacy budget simply because you're going to be using those data points only once every so many days.

**WALDO:** It seems that many of the examples you're offering are gross in the sense that there are fairly large numbers of entities in the datasets on one side or the other of a comparison—meaning that adding a small amount of noise really shouldn't cause an issue. But I wonder about queries around outliers. Say, if I wanted to find the number of people in some particular country who were still running Windows XP or maybe were still using OS/2, a little bias in those numbers would probably cause a real difference in the outcomes. When do you think it's appropriate—or inappropriate—to use a query that is differentially private?

**GUEVARA:** In general, I think differential privacy is very good for describing broad statistical trends in terms of how thousands of people do X things each day. The Community Mobility Reports that Google has been producing to track COVID-19 infection trends is a good example. There are other use cases where you can look at some very particular abuse or spam trends indicating specific attack vectors. If you end up doing some very granular queries on that, you'll find that—while it's theoretically possible to accomplish this with differential privacy—the relative impact of the noise will be so huge that the results you'll get will be almost useless.

As a general rule, I'd say that while differential privacy is good for doing broad population analysis, it's not so good at figuring out how one or two people are behaving since, by definition, that's the very thing differential privacy is designed to protect against.

**COATTA:** A couple of times already we've made reference to the amount of



MIGUEL GUEVARA

**As a general rule, I'd say that while differential privacy is good for doing broad population analysis, it's not so good at figuring out how one or two people are behaving since, by definition, that's the very thing differential privacy is designed to protect against.**





JIM WALDO

**One of the interesting things I've observed about differential privacy is that there has been about a 10-year lag between the theoretical foundations and the first practical applications, which are only now becoming available.**



privacy that might be “lost,” whereas the layman concept of privacy is more Boolean—that is, it's either private or not. So, it's interesting to talk about it here as a quantitative measure. What does that actually mean?

**DESFONTAINES:** The notion of privacy as something Boolean is misleading from the start. Even outside of differential privacy, you always need to ask yourself questions like: How can we make this feature work while collecting as little data as possible? What level of protection should we apply to the data we store? How can we request user consent in an understandable, respectful way? And so on.

None of these questions is Boolean. Even in adversarial contexts, where the answer *seems* to be Boolean, it isn't. For example: Is the attacker going to be able to intercept and re-identify data? The answer is either yes or no.

But you still need to think about other questions like: What is the attacker capable of? What are we trying to defend against? What's the worst-case scenario? This is to say, even without the formal concept of differential privacy, the notion of privacy in general is far from Boolean. There always are shades of gray.

What differential privacy does to achieve data anonymization is to quantify the trade-offs in a formal, mathematical way. This makes it possible to move beyond these shades-of-gray assessments to apply a strong attack model where you have an attacker armed with arbitrary background knowledge and computational resources—which represents the worst possible case—and yet you're still able to get strong, quantifiable guarantees. That's the essence of differential privacy, and it's by far the best thing we have right now in terms of quantifying and measuring privacy against utility for data anonymization.

Powerful as differential privacy may be, it's also highly abstract. Getting users and developers alike to build confidence around its ability to protect personally identifiable information has proved to be challenging.

In an ongoing effort, various approaches are being tried to help people make the connection between the

mathematics of differential privacy and the realization of actual privacy protection goals. Progress in this regard is not yet up to Google scale.

And yet, Google has a clear, vested interest in building public confidence in the notion that it and other large aggregators of user data are fully capable of provably anonymizing the data they utilize. Finding a way to convey that in a convincing manner to the general public, however, remains an unsolved problem.

**WALDO:** When it comes to users who are worried about privacy, I doubt you'll be able to ease those concerns much by telling them you've set epsilon to some particular value. How do you translate the significance of that into something users can understand?

**GUEVARA:** Honestly, I don't think we've done a great job of communicating this to users. We've been more focused on raising awareness. But this issue you raise is an important one since there are just so many misconceptions about anonymization out in the world right now. Many people believe that, to anonymize data, you just remove an entire identifier from a dataset. So, our first step is to make sure everyone realizes that does *not* qualify as proper or strong anonymization.

Then, once we get to that stage where users have personal privacy as their mindset, one of the biggest priorities for those of us in the privacy research community needs to become exactly what you're saying: How can we help people see the connection between what we're doing mathematically and what they've come to expect in terms of protections for their own personal privacy?

We've already done a bit of user research that has allowed us to really talk with users, and what I've learned is that, whenever we're able to show people how their personal data can be hidden behind the crowd and protected by random information, they definitely come around to expressing more confidence. Clearly, however, there's still a huge challenge ahead for us in terms of learning how to talk about these mathematical techniques and the guarantees they confer in ways that feel more tangible to end users.



**DESFONTAINES:** The other side of this is that gaining a better understanding of the users' privacy concerns is part of what informs policy. Some of their questions are entirely orthogonal to the use of differential privacy. For example: Who among my family and friends and colleagues can see what I just shared online? How long will my data be kept?

When the time comes, we need to be able to offer differential privacy as an answer to the different, more specific question: How is my data protected whenever Google shares aggregated data publicly?

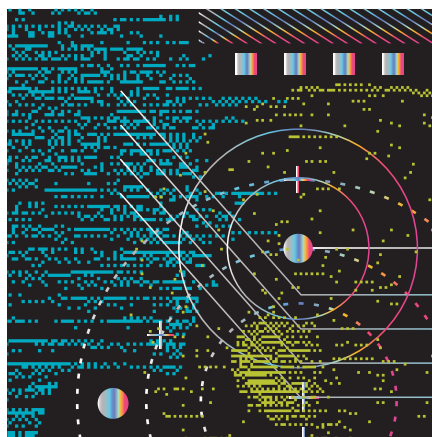
**WALDO:** Maybe you ought to describe what you've developed for Google to make differential privacy a little easier for the average programmer to use.

**GUEVARA:** The first critical thing to point out is that we've developed a SQL engine that produces differential privacy results. The core idea behind that was, since a lot of analysts are already familiar with SQL, it would be best just to augment that syntax with a couple of differentially private operations. Essentially, that means someone can do an anon count and produce a differential privacy count from that and, similarly, do an anon sum and produce a differential privacy sum.

Some of the other pieces we've built are geared more toward a data-operation framework that processes a lot of data. You can think of them as Apache Beam-type frameworks that let us turn regular operations—primarily counts and sums—into differential privacy operations that teams then can use to produce their data in a manner that better protects privacy. [Apache Beam is an open source, unified model for defining both batch and streaming data-parallel processing pipelines.]

**WALDO:** How broadly is this used within Google and in what context?

**DESFONTAINES:** Probably the most visible user-facing examples are a few features in Google Maps that are powered by differential privacy. Then there also are the COVID-19 Community Mobility Reports mentioned earlier. We use differential privacy internally as well to help analysts access data in a safe, anonymized way, and to power internal dashboards that let developers monitor how their products are being used. Basically, at a high level, any time a



team wants to do something with sensitive data that calls for the data to be handled in an anonymous manner—for example, to retain the data longer so that data-protection requirements that might otherwise call for encryption or tight access controls can instead be relaxed—we encourage them to use differential privacy.

**COATTA:** But I can easily imagine users shooting themselves in the foot when using differential privacy. For example, I might issue some queries against the database, get some results back, and think I actually know what those results mean. What I might fail to recognize is that there's so much noise in those results that they actually don't mean anything at all. What does Google's differential privacy library do to help people avoid this trap?

**GUEVARA:** Results that contain more noise than you realize can be a real problem from a usability perspective. In fact, one of the things our internal users continually ask us is: Where should we stop trusting the data?

Imagine that you issue a query and then get back a table that, say, gives you different counts. At some point, those counts will have more noise than real data in them. One way we try to address that is by providing confidence intervals in the results, with the hypothesis being that, if the confidence interval is very small relative to the value, then there's very little noise—meaning users can trust that result. If the confidence interval is very broad, then users can infer there's a lot of noise. And then, yes, they can stop trusting the data at that point.

**DESFONTAINES:** In the specific use case of the COVID-19 Community Mobility Reports, which contain data that



TERRY COATTA

**I can easily imagine users shooting themselves in the foot when using differential privacy. For example, I might issue some queries against the database, get some results back, and think I actually know what those results mean.**



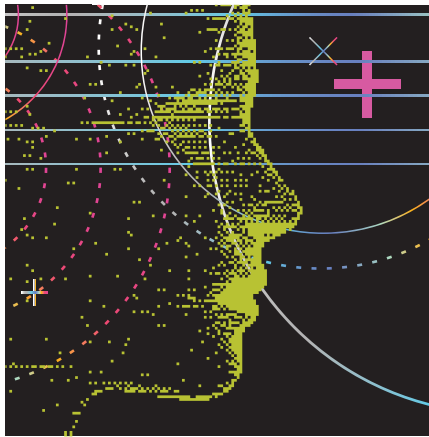
researchers and policymakers use to make hard decisions about social distancing and that sort of thing, we don't want them to derive the wrong conclusions from the data just because they don't really understand the noise-addition process. We did a couple of things to help avoid that. One is that we decided to publish only the data where the confidence intervals seemed tight enough. That is, if the added noise had more than a 10% chance of leading to numbers that were more than 10% off, we didn't release that data. Instead, we'd say, "What we have isn't accurate enough, so no data is available for this metric."

The second thing we did was to document the whole process just as precisely as we could in a whitepaper that's been published online [Differentially Private SQL with Bounded User Contributions; <https://arxiv.org/abs/1909.01917>]. Referring to this, anyone doing complex statistical analyses on the data should have what they need to account for the uncertainty contributed by the noise.

**WALDO:** Of course, any machine-learning algorithm also has a certain confidence interval. What is the relationship between the confidence intervals you're able to get out of a differentially private query on the data and what the machine-learning folks then manage to do with that data? Or have you not connected the two as yet?

**DESFONTAINES:** There are various ways to combine differential privacy and machine learning, and we have a lot of researchers working on that very thing—in particular, by increasing the accuracy of machine-learning models while making them safe through the use of differential privacy. We've also published an open source library [TensorFlow Privacy] that incorporates some of these techniques as part of training models for machine learning.

We're now experimenting to understand better how machine-learning models trained on sensitive datasets can inadvertently memorize information from the original training data, while also working to see how differential privacy might be used to quantify that. One challenge is that the epsilon parameters we get by way of these methods are typically quite high,



sometimes to the point where it's hard to interpret the relevant guarantees. Empirically, however, it also seems that even these difficult-to-interpret guarantees generally prove successful in mitigating attacks. Let's just say this is proving to be a fascinating and fruitful field of research.

**COATTA:** Have you run into any complications in trying to combine differential privacy with other privacy-protecting technologies? I ask, since differential privacy clearly isn't going to solve all our problems.

**GUEVARA:** I think we're just too early in our efforts to advance protections to know what all the possibilities are, but there are some encouraging signs. I've heard that some people are trying to use differential privacy with federated learning to train models in a provably private way. I've also heard that differential privacy is being used together with homomorphic encryption to share data between two parties such that both parties then can produce results that don't reveal any individual patterns or any group of patterns.

**WALDO:** One of the interesting things I've observed about differential privacy is that there has been about a 10-year lag between the theoretical foundations and the first practical applications, which are only now becoming available. What has made this so difficult to implement?

**DESFONTAINES:** We were quite surprised by some of the difficulties we encountered. Fundamentally, I don't think the math is all that hard. The basic results and techniques are relatively simple, and it doesn't really take much time or effort to get a reasonable understanding of the theory behind them. But it turns out that transforming all

that theory into practice has proved difficult and has required more time and thought than we anticipated.

There are a few reasons for this. One is that the literature makes some assumptions about the type of data you'd be looking to anonymize, and we discovered—in practice—that this is mostly wrong. An example is the assumption that each record of the dataset corresponds to a single user. This owes to the fact that the main use case presented in much of the literature relates to medical data—with one record per patient. But, of course, when you're working with datasets like logs of user activities, place visits, or search queries, each user ends up contributing much more than just a single record in the dataset. So, it took some innovations and optimizations to account for this in building some better tooling for our purposes.

Something else that contributed to the unforeseen difficulties was that, even though the math is relatively simple, implementing it in a way that preserves the guarantees is tricky. It's a bit like RSA (Rivest-Shamir-Adleman) in cryptography—simple to understand, yet naïve implementations will encounter serious issues like timing attacks. In differential privacy theory, you add a random number from a continuous distribution to a statistic with arbitrary precision. To do that with a computer, you need to use floating-point numbers, and the ways these are represented come with a lot of subtle issues. For example, the bits of least precision in the noisy number can leak information about the original number if you're not careful.

---

In many ways, the release of an open source version of Google's differential privacy library creates a whole raft of new challenges. Now there's an education program to roll out; users and developers to be supported; new tools to be built; external contributions to be curated, vetted, and tested ... indeed, a whole new review process to put into place and an even broader undertaking to tackle in the form of organizing an external community of developers.

But that's just what comes with the territory whenever there are grand aspirations. The goals of Google's differential privacy team happen to be quite ambitious indeed.

**COATTA:** It's great you've released this open source library that provides for the implementation of much of the really subtle mathematical computation at the heart of differential privacy. But do you also have a lot of unit tests to make sure this isn't going to go off the rails?

**DESFONTAINES:** One of the other things we open sourced along with the library was a testing framework, specifically built to verify differential privacy guarantees. But unit tests are a little difficult for that type of library. By design, differential privacy randomizes its outputs, so you can't simply check to make sure the value returned is the one you were expecting. The testing framework, on the other hand, gives you a way to empirically verify the formal property of differential privacy by generating lots of outputs and applying statistical tests. We published a description of one of our methods in the whitepaper I referred to earlier.

Anyway, yes, we agree: Testing is super-important, and special statistical techniques must be used to complement unit testing and manual auditing.

**WALDO:** In looking over your open source page, I see a few languages are supported—one seemingly better than the others. Do you plan to expand this to other languages? Or are you going to focus more on adding new algorithms? Do you think you'll manage to do a bit of both?

**GUEVARA:** The languages supported are those we use in production at Google: Go, C, and Java. In time, we hope to offer the same set of features for each of those three languages. You'll probably soon see an experimental folder that will contain some new things like those higher-level, data-processing frameworks I mentioned earlier. There also will be some open source things to help with the accounting for privacy budgets over a set of queries. We're definitely looking to extend our open source library, and the things people will find there are mostly the same things we use internally, meaning we have a lot of confidence in them.

**COATTA:** What if people outside of Google encounter difficulties when using the technology? After all, it's not as if they can walk down the hall to talk to the person who wrote whatever it is they're having an issue with.

**GUEVARA:** We try to answer people's questions on the repository to the degree possible. Anyone can check the comments posted there and the issues submitted there. Our goal, actually, is to be as supportive as possible.

**WALDO:** It looks at this point as though this is mostly a read-only open source repository for people outside of Google. Do you have any plans to expand the implementation team to include people from outside?

**DESFONTAINES:** In time we'd like to open it up to external contributions. At first, our C++ library didn't seem to generate a lot of external contributor interest. For one thing, the number of people who work on differential privacy isn't huge, and C++ isn't widely used in the open source community. Still, more recently, we've witnessed a real growth in interest, both for differential privacy in general and for our work in particular. Folks at OpenMined, for example, wrote a Python wrapper for our work and are working on Java tooling based on our libraries. We hope to attract even more people as we start to publish more in Java and Go—in particular around end-to-end tools like Privacy on Beam.

**WALDO:** Whenever the time comes for you to start taking in external contributions, it should make for an interesting vetting process since this is fairly subtle stuff.

**DESFONTAINES:** Exactly. Much remains to be determined in terms of what we'll need to do in the way of testing, mathematical proofs, ensuring code quality, and the rest of it prior to accepting any contribution into the repository.

**GUEVARA:** We'll need to make sure the differential privacy mechanisms are actually doing what they're supposed to be doing—which means there would need to be some sort of review process. We're just not sure what that process ought to look like yet.

**COATTA:** How widely deployed do you expect differential privacy ultimately to be?

**GUEVARA:** It could have the sort of reach encryption has currently. In the same way that many people now use

encryption by default, I'd like to see a world where people use differential privacy by default prior to analyzing datasets. That should just be a standard best practice. That's because privacy protections then would become commonplace.

There's another aspect of differential privacy we haven't talked about yet, and that's the ability to collect data in a differentially private way. So, here again, going back to that crypto analogy, I'd like to see a world where, by default, data applications collect data *only* in a differentially private manner—perhaps allowing exceptions only for specific use cases.

**DESFONTAINES:** I agree with Miguel. The biggest barrier to achieving differential privacy today is not the math or a lack of theoretical research. Instead, we need more implementations and some dedicated effort to make differential privacy easier to use. Once we have that, people will be able to readily add differential privacy whenever they're publishing the results of data analysis or statistical studies. Then, maybe differential privacy will become a standard best practice rather than just a curiosity.

Should similar efforts around local differential privacy also prove successful, that too could become a best practice for data collection—at least, that's a long-term goal of ours. The only thing that stands between us and achieving that goal is more implementation, usability, and outreach work—as opposed to more research breakthroughs.

**COATTA:** In terms of this becoming the default way of doing data analysis, how long will it be before a differentially private data service becomes something I can just sign up for on the Google engine or AWS (Amazon Web Services)?

**GUEVARA:** A lot of the foundational pieces already exist on the Google site, so I don't think it should take that long. My optimistic estimate would be one year. A pessimistic estimate would be more like three years. But I sure hope it doesn't take that long before we're able to offer default services that deliver differential privacy for end users in a more intuitive manner. □



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

## Why DevOps encourages us to celebrate outages.

BY THOMAS A. LIMONCELLI

# The Time I Stole \$10,000 from Bell Labs

IF IT WORKERS fear they will be punished for outages, they will adopt behavior that leads to even larger outages. Instead, we should celebrate our outages: Document them blamelessly, discuss what we've learned from them openly, and spread that knowledge generously. An outage is not an expense. It is an investment in the people who have learned from it. We can maximize that investment through management practices that maximize learning for those involved and by spreading that knowledge across the organization. Managed correctly, every outage makes the organization smarter. In short, the goal should be to create a learning culture—one that seeks to make only new mistakes.

I worked at Bell Labs in New Jersey from 1994 to 2000. I was a systems administrator on a team of people charged with maintaining thousands of computers

and the network that connected them. It was intimidating to be surrounded by so many brilliant scientists and engineers, many of whom had written the textbooks I used in college.

One day, I had to make a configuration change to the central router. It is difficult to measure the size of a change. I could say it was a tiny change in that it affected only a few lines of the router's configuration file. On the other hand, it was a big change in that it impacted a network used by thousands of users. It was an important change because an important project was blocked waiting for it to be completed.

I typed the commands to alter the configuration, saved the new configuration, and checked the things I usually check. The change was a success ... or so I thought.

Proud of myself, I moved on to other work. A little while later I couldn't connect to most machines on the network. Neither could anyone else. I panicked. Could my change have caused that? Impossible! That was nearly an hour ago.

No, it was definitely my change. There are some typos that don't show any ill effects right away. In this case, a cache was held for 45 minutes. At 46 minutes the router was a very expensive box doing nothing.

I reverted my change, and everything returned to normal.

My father used to joke about weather forecasters. For example, he would say that if they simply predicted that tomorrow the weather would be "about the same as today," they would be accurate 70% of the time where we lived, and perhaps 90% of the time in Los Angeles. By way of analogy, I often joke that during an outage, asking, "What was the last big change we made?" will make you look like a genius 70% of the time, and perhaps 90% of the time in Los Angeles.

Even though my change had been completed nearly an hour ago, it was certainly the most recent big change.

### Learning The Wrong Lesson

Sitting at my desk, I did a little back-of-the-envelope math to calculate the cost



of this outage: number of people affected, estimated average Bell Labs salary, the likelihood people were at their computers at the time...

My calculations estimated that the outage cost the company about \$10,000 and affected thousands of people.

I panicked.

I hid in my office.

Prayed that nobody would say anything or notice.

And guess what? Nobody did.

I dodged the bullet. Or, maybe someone else was blamed. I didn't care as long as I didn't get in trouble.

I learned an important lesson that day: Don't touch that particular configuration parameter and, if you do, always wait at least an hour before you declare success.

### **And Then It Happened**

It may surprise you, but that outage was not the time I stole \$10,000 from Bell Labs.

It was an honest mistake, a beginner's mistake. Chalk it up to the cost of learning on the job. While my fear and embarrassment were real, most likely those feelings were unfounded. I had an

awesome boss who would have protected me. Plus, LANs were pretty unreliable back then, and most of the affected users probably took the outage in stride.

The stealing was what happened next.

A month later another person on my team made the exact same mistake. The outage was the same size, duration, and estimated \$10,000 cost.

That outage definitely would have been prevented if I had shared what I learned from *my* outage. I knew it then, and years later I still believe it.

The stealing wasn't a result of my outage. It was how I responded to my outage. I had robbed the company of the opportunity to learn and improve.

### **Fear Drives Negative Behavior**

If people are afraid they will be punished for outages, the result will be self-protective behaviors that have unintended negative side effects. These side effects can lead to more frequent and bigger outages.

Some of these negative behaviors include:

► *Hiding mistakes.* This blocks organizational learning and can rob the company of potential improvements.

► *Hiding problems.* People will intentionally hide a problem if there is a culture of shooting the messenger. This leads to problems being discovered only when they are too big to be invisible.

► *Ignoring small problems.* People will ignore a small problem out of fear that fixing it, which is often error prone, may lead to an outage that they will be blamed for. This leads to problems being addressed only when they are big enough, and expensive enough, that they can't be ignored.

► *Shutting down communication.* Fear has a chilling effect that prevents the open and honest communication required to work well as a team, and prevents teams from working well together.

► *Losing the best skilled people.* If they don't choose to leave the toxic culture, the toxic culture will force them out.

If your organization runs from one major disaster to another, maybe the problem is a corporate culture that unintentionally drives these behaviors.

Want a more reliable system? You will need a team that is highly skilled, communicates effectively, and fixes problems when they are small. Fear creates the opposite.

Frequently after a major outage or other problem, we see CEOs or politicians claiming they will “fire the responsible person.” Congrats, dude. You just helped assure a future full of bigger, more frequent outages.

I’m not sure where this “fire someone” response comes from. It certainly makes good TV sitcom material. It definitely plays well at a news conference. It’s doubtful, however, that MBA programs are teaching future executives that if they fire anyone who makes a mistake, eventually their company will employ only perfect people. On the contrary, firing everyone who makes a mistake will result in a company with no employees, or a company full of people waiting to be fired when management discovers that they are human. Yet, frequently CEOs and politicians are pressured to prove their seriousness by firing someone. How many times did pundits speculate when or who President Obama would fire during the stunted launch of the Affordable Care Act website?

Such toxic cultures make it difficult to hire the best. Word travels fast. If your company has a reputation for blaming and shaming, word will spread, and top talent will avoid you.

### DevOps Celebrates Mistakes

DevOps culture has a more enlightened attitude about outages. Rather than hiding them or pretending they didn’t happen, we document them. Rather than punishing anyone, we encourage responsibility and accountability.

It is irrational to believe that a complex system can be 100% free of outages. Therefore, punishing people or getting angry at someone because of an outage is irrational.

A more enlightened stance is to view each outage as an unplanned investment. I didn’t create a \$10,000 outage at Bell Labs. Bell Labs invested \$10,000 in my education. To make the most out of that investment, the education should be put to the best use possible.

Learning from incidents does not magically happen. The desire may exist, but more is required. The shift from blame to learning requires a commitment from executives, management, and non-management alike. Executives must model blameless behavior and encourage learning. Management

must create processes that enable learning. Project managers need to allocate space and time for these processes to happen. Everyone must learn to be more open and humble.

DevOps culture encourages writing a postmortem report to capture what happened and what was learned. Focusing on the question, “What was learned?” rather than, “Why did this happen?” or “Who’s to blame?” creates a culture of learning and improvement.

Postmortems help us to be accountable. The word *accountable* literally means “to account for what happened”—that is, to tell the story. The postmortem should focus on a timeline of what happened and what was learned.

A postmortem report usually concludes with a list of what should be done to prevent similar events in the future. Each item in that bullet list is triaged like any other bug or feature request. New thinking in DevOps suggests that focusing on this list is a distraction from the learning process. Some organizations have started to separate out the process of identifying these follow-up projects by moving that discussion to a separate meeting conducted afterwards, often with a smaller group of people.

Dave Zwieback’s excellent book *Beyond Blame: Learning from Failure and Success* discourages the use of the term *postmortem* and instead calls the process a *learning review*. A learning review can be used to analyze any event. There is as much to learn from success as from failure.

Large events (outages and successes) are chock-full of learning opportunities. Those involved should be encouraged to share what was learned even more widely by presenting on the topic. While at Google, I frequently saw SREs (site reliability engineers) travel to far-flung offices to give presentations about a recent outage and how the local team could leverage what was learned. Talk about the opposite of hiding in shame!

When an outage affects customers, a public version of the postmortem report should be made available. Public relations and legal departments will likely break into a sweat the first time this is suggested, but companies are learning that public postmortems actually build customer confidence and loyalty.

The best public postmortems present what has been learned in ways that are useful to customers. The highest compliment you can get is, “I learned so much from your public postmortem that it made me better at my job!” What customers mean by such a compliment is that either they have learned a practice they can adopt at their company, or they have learned previously obscured details about your product that help them do their jobs better when using your product. The loyalty this creates is priceless.

It is important that communication with the public be authentic. Sound like a human, not a press agent. Admit failure. Write in the first person and show real remorse. Avoid the temptation to minimize the full impact of the outage by saying, “We regret the impact it may have had on our users and customers.” *May* have had an impact? There *was* impact! Otherwise, you wouldn’t be sending this message. Say, “We apologize for the impact this outage had on our customers.” Your legal and public relations departments may have trouble with this at first, but they need to learn that today’s customers are astute judges of authenticity.

### Conclusion

Obviously, I didn’t literally steal \$10,000 from Bell Labs. But I did rob my team of learning from my mistake in a way that could have improved the entire team. I learned my lesson, and I’m glad to have the opportunity to share it with you.

Nobody loves outages. They are inevitable, so we might as well make the most of them. Through blameless postmortems and other techniques we can create a culture where every outage results in the organization becoming smarter.

If we do it right, the only mistakes we make will be new mistakes. ■

---

*If you would like to learn more about this subject, I recommend Zwieback’s Beyond Blame: Learning from Failure and Success and chapter 14 of The Practice of Cloud System Administration, the book I wrote with Strata Chalup and Christina Hogan.*

---

**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 and tweets at @YesThatTom.

Copyright held by author/owner.  
Publication rights licensed to ACM.



# 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

DOI:10.1145/3382035

**A new French keyboard standard is the first designed with the help of computational methods.**

BY ANNA MARIA FEIT, MATHIEU NANCEL, MAXIMILIAN JOHN, ANDREAS KARRENBauer, DARYL WEIR, AND ANTTI OULASVIRTA

# AZERTY amélioré: Computational Design on a National Scale

IN 2015, FRANCE'S Ministry of Culture wrote to the French Parliament<sup>4</sup> criticizing the lack of standards for a keyboard layout. It pointed out that AZERTY, the traditional layout, lacks special characters needed for “proper” French and that many variants exist. The national organization for standardization, AFNOR, was tasked with producing a standard.<sup>5</sup> We joined this project in 2016 as experts in text entry and optimization.

THE FRENCH LANGUAGE uses accents (for example, é, à, î), ligatures (œ and æ), and specific apostrophes and quotation marks (for example, ‘ « » “ ”). Some are awkward to reach or even unavailable with AZERTY (Figure 1), and many characters used in French dialects

are unsupported. Similar-looking characters can be used in place of some missing ones, as with “ for “, or ae for œ. Users often rely on software-driven auto-completion or autocorrection for these. Also, they insert rarely used characters via Alt codes, from menus, or by copy-pasting from elsewhere. The ministry was concerned that this hinders proper use of the language. For example, some French people were taught, incorrectly, that accents for capital letters (for example, É, À) are optional, a belief sometimes justified by reference to their absence from AZERTY.

This article reports experiences and insights from a national-scale effort at redesigning and standardizing the spe-







ILLUSTRATION BY MATT HERRING

cial-character layout of AZERTY with the aid of *combinatorial optimization*. Coming from computer science, our starting point was the known formulation of keyboard design as a classical optimization problem,<sup>2</sup> although no computationally designed keyboard thus far has been adopted as a nationwide standard. The specific design task is shown in Figure 2. Going beyond prior work, our goal was not only to ensure high typing performance but also to consider ergonomics and learnability factors.

However, the typical “one-shot” view of optimization, in which a user defines a problem and selects a solver, offers poor support for such complex

socio-technical endeavors. The goals and decisions evolved considerably throughout the three-year project. Many stakeholders were involved, with various fields of expertise, and the public was consulted.<sup>19,20</sup> A key takeaway from this case is that algorithmic methods must operate in an interactive, iterative, and participatory manner, aiding in defining, exploring, deciding, and finalizing the design in a multi-stakeholder project.

In this article, we discuss how interactive tools were used to find a jointly agreed definition of what makes a good keyboard layout: familiarity versus user performance, expanded character sets versus discoverability, and support for

## » key insights

- France is the first country in the world to adopt a keyboard standard informed by computational methods, improving the performance, ergonomics, and intuitiveness of the keyboard while enabling input of many more characters.
- We describe a human-centric approach developed jointly with stakeholders to utilize computational methods in the decision process not only to solve a well-defined problem but also to understand the design requirements, to inform subjective views, or to communicate the outcomes.
- To be more broadly useful, research must develop computational methods that can be used in a participatory and inclusive fashion respecting the different needs and roles of stakeholders.






and experts in ergonomics, typography, human-computer interaction, linguistics, and keyboard manufacturing. A typical standardization process involves meetings to iterate over each aspect of the standard and its wording. Final drafts are opened to public comment on which the committee then iterates if need be. At the start of the project, we took these meetings as an opportunity to understand the requirements of the design problem from a human-centered perspective. We then formulated them in a way that enabled modeling and solving the problem using optimization.


Our task was to develop an improved layout for all so-called “special characters”, that is, every character that is not a nonaccented letter of the Latin alphabet (“AZERTYUIOP...”), a digit, or the space bar. The list of special characters to be made accessible was greatly augmented compared to the traditional AZERTY layout, to facilitate the typing of all characters used in the French language and its dialects,<sup>a</sup> modern computer use (especially programming and social media), and scientific and mathematical characters (for example, Greek letters), alongside major currency symbols and all characters in Europe’s other Latin-alphabet languages. Despite having to add many new characters, we strove to keep the layout usable, ergonomic, and easy to learn.

There were several challenging requirements (Figure 2). The physical layout follows the alphanumeric section of the ISO/IEC 9995-112 standard. Each key can hold up to four characters, using combinations of the Shift and AltGr modifiers (Figure 2c). For nonaccented letters, digits, and the space, the layout had to remain as in traditional azerty, leaving 129 keyslots (see Figure 2b). The only characters that could be added or moved were the special characters described in Figure 2a; their number, up to 122, varied throughout the project as new suggestions were made and discussed. Combining diacritical characters, like accents, are entered via “dead keys,” as explained in Figure 2d.

Note that the requirements and constraints of this project evolved dramatically as it progressed, depending on



**Despite having to add many new characters, we strove to keep the layout usable, ergonomic, and easy to use.**



intermediate solutions, priorities updates, public requests, and so on. We detail these changes in the later text.

### **Keyboard Design as an Optimization Problem**

The arrangement of characters in a layout is a very challenging computational problem. Formally, one must assign characters to the keyboard keys and to keyslots accessible via modifier keys. Each assignment involves three challenging considerations. We here discuss the computational problem before opening up approaches to making them useful in a multi-stakeholder design project.

Firstly, what is a “good” placement? Ergonomics and motor performance should be central goals. More common characters should be assigned to keys that minimize risks of health issues such as repetitive strain injury and that are quickly accessed. However, people differ in how they type.<sup>7</sup> There is no standard model that can be used as an objective function. Also, time spent visually seeking a character should be minimized through, for example, placing characters where people assume they are,<sup>14</sup> and grouping characters that are considered similar.

Secondly, which level of language to favor is tricky to know in advance and, as we learned, a politically loaded question. To decide where to put #, we must weigh the importance of programming or social-media-type language in which that character might be common, against “proper” literary French in which it is rare. Decisions on character positions mean trading off many such factors for a large range of users and typing tasks.

Finally, there is a very large number of possible designs, up to  $10^{213}$  distinct combinations for assigning characters to keyslots in our case. Text input is a sequential process wherein entering a character depends on the previously typed one. Therefore, finding the best layout for typing is an instance of the quadratic assignment problem (QAP).<sup>2,6,16</sup> These are not only hard to solve in theory (NP-hard to approximate within any constant factor<sup>22</sup>); there still exist unsolved instances of QAPs, published as benchmarks decades ago, with only 30 items,<sup>3</sup> a far cry from 129.

<sup>a</sup> <https://bit.ly/32ZGnQh>

**An optimization model for typing special characters.** The design problem was formulated as an integer program (IP), which lets us use effective solvers that provide intermediate solutions with bounds on their distance from optimality. We use binary decision variables  $x_{ik}$  to capture whether character  $i$  is assigned to keyslot  $k$  or not. The criteria, and corresponding IP constraints, are formulated in Table 1. Every feasible binary solution corresponds to a keyboard layout. An objective function measures the goodness of each layout according to each of the criteria. The parameters, constraints, and objectives of the integer program reflect the standardization committee’s goals: facilitate typing of correct French, enable the input of certain characters not supported by the current keyboard, and minimize learning time by guaranteeing an intuitive to use keyboard that is sufficiently similar to the previous AZERTY.

The challenge for us was to translate goals such as “facilitate typing and learning” into quantifiable objective functions. We ended up defining four objective criteria, which were combined in a weighted sum to yield a single objective function.<sup>6,19</sup> **Performance** (minimizing movement time), **Ergonomics** (minimizing risks of strain), **Intuitiveness** (grouping similar characters together), and **Familiarity** (minimizing differences from AZERTY). Table 1 presents our formulation of the integer program and articulates the intuition behind each criterion.

The criteria here rely on input data that reflect the real-world typing of tens of millions of French users. Therefore, we gathered large text corpora, with varied topics and writing styles, and weighted them in accordance with the committee’s requests. We focused on three typical uses. Formal text is well-written, curated text with correct French and proper use of

special characters. Sources include the French Wikipedia, official policy documents, and professionally transcribed radio shows. Informal text (for example, in social-media or personal communication) has lower standards of orthographic, grammar, and typographic correctness. The material includes anonymized email and popular accounts’ Facebook posts and Tweets. The Programming corpora comprise content representative of common programming and description languages: Python, C++, Java, JavaScript, HTML, and CSS, with comments removed. Several of our Formal-and Popular-class corpora were provided by the ELDA.<sup>8,b</sup> Frequencies were computed by corpus, then averaged per character and class, and finally assigned weights subject to committee discussion (Formal: 0.7, In-

b The Evaluations and Language resources Distribution Agency; see <http://www.elra.info/en/about/elda/>.

**Table 1. The integer programming formulation of the keyboard design problem. The objective function is a weighted sum over four normalized criteria. Only basic assignment constraints are shown. Throughout the project, additional constraints were added or removed for particular instances.<sup>8</sup> For the instance that led to the standardized layout ( $N = 85$ ,  $M = 129$ ), the following weights were chosen:  $w_P = 0.3$ ,  $w_E = 0.25$ ,  $w_I = 0.35$ ,  $w_F = 0.1$ .**

$\min \quad w_P \sum_{i=1}^N \sum_{k=1}^M \sum_{c=1}^{27} (p_{c'} T_{ck'} + p_{ic} T_{kc}) x_{ik}$ $+ \quad w_E \sum_{i=1}^N \sum_{k=1}^M p_i (W_k + F_k + M_k) x_{ik}$ $+ \quad w_I \left( \sum_{i=1}^N \sum_{j=1}^N \sum_{k=1}^M \sum_{l=1}^M S_{ij} D_{kl} x_{ik} x_{jl} + \sum_{i=1}^N \sum_{k=1}^M \sum_{c=1}^{27} S_{ic} D_{kc} x_{ik} \right)$ $+ \quad w_F \sum_{i=1}^N \sum_{k=1}^M p_i D_{kA(i)} x_{ik}$	<p><b>Performance (P):</b> Guarantees that frequent special characters can be quickly entered in combination with the fixed letters. It is quantified by computing the average time to type a special character before or after any of the regular letters (<math>T_{ck'}</math>, <math>T_{kc}</math>), weighted by the special-character–regular-letter pair (<math>p_{c'}</math>, <math>p_{ic}</math>). The corresponding data were gathered in a crowdsourcing-based study.</p> <p><b>Ergonomics (E):</b> Penalizes keyslots that require extreme movements putting strain on tendons and joints, which are empirically associated with repetitive strain injuries<sup>24</sup>: extreme outward or inward movements of the wrist (<math>W_k \in [0, 1]</math>), extreme extension of fingers (<math>F_k \in [0, 1]</math>), and use of one or two modifier keys (<math>M_k \in [0, 1, 2]</math>). The score is weighted by the frequency (<math>p_i</math>) of the character assigned to the keyslot.</p> <p><b>Intuitiveness (I):</b> Minimizes the distance between similar special characters (<math>D_{kl}</math>) and between special characters and similar letters (<math>D_{kc}</math>), to facilitate discovery and learning.<sup>17</sup> This similarity can be syntactic or semantic and is captured by the scores <math>s_{ij}</math>, <math>s_{ic}</math>. All characters are considered equally important for grouping.</p> <p><b>Familiarity (F):</b> Places frequent characters near the position familiar from traditional AZERTY, to facilitate visual search with the new layout.<sup>14</sup> <math>D_{kA(i)}</math> quantifies the distance between the keyslot <math>k</math> assigned to the character <math>i</math> and its AZERTY position, weighted by that character’s frequency (<math>p_i</math>).</p>
<p>subject to</p> $\sum_{k=1}^M x_{ik} = 1 \quad \forall i \in [1, \dots, N]$ $\sum_{i=1}^N x_{ik} \leq 1 \quad \forall k \in [1, \dots, M]$ $x_{ik} \in [0, 1] \quad \forall i \in [1, \dots, N], k \in [1, \dots, M]$	<p>Ensures each character is assigned to one keyslot.</p> <p>Ensures no keyslot is assigned to multiple characters.</p>



formal: 0.15, Programming: 0.15). Table 2 shows the most common characters in each category.

For estimating key-selection times, we gathered an extensive dataset of key-to-key typing durations to capture how people type in terms of the Performance objective. In particular, we were interested in how soon a special character keyslot (in green in Figure 2) could be accessed before or after a regular letter. In a crowdsourcing-based study, we asked about 900 participants to type word-like sequences of nonaccented letters that each had one special character slot in the middle,<sup>6,19</sup> for example, “buve Alt+Shift+2 ihup.” We gathered time data for all combinations of letters and special character slots (7560 distinct key pairs).

For the Intuitiveness objective, we defined a similarity score between characters as a scalar in the range [0, 1], depending on visual resemblance (for example, R and @, \_ and -), semantic proximity (for example, × and \*, or ÷ and /), inclusion of other letters (for example, ç and c, œ and o), frequent association in practice (for example, n and ~, e and ’), or use-based criteria such as lowercase/uppercase and opening/closing character pairs. These weights, and the similarities to consider and give priority, were discussed at length with the committee and frequently updated throughout the project, especially after the public comment.

**Solving the QAP.** Branch-and-bound<sup>1</sup> is a standard approach to solve integer optimization problems. It relies on relaxations that can be solved efficiently (for example, by linearizing the quadratic terms and dropping the integrality constraints). In the powerful RLT1 approach,<sup>10</sup> every quadratic term of the form  $x_{ik} \cdot x_{kl}$  is replaced with a new linear variable  $y_{ijk}$ . Although this linearization produces very good lower bounds, it introduces  $\mathcal{O}(n^4)$  additional variables, leading to a vast increase in problem size. We observed, however, that, although we have 100+ characters to place, our quadratic form is very sparse. Our approach exploits this sparsity, leading to a framework that synthesizes the concepts and benefits of powerful (but complex) linearization and column-generation technique.<sup>13</sup>

In our adaptation, only a subset of variables is part of the initial instance, and further variables are generated

**Table 2. The highest-frequency special characters, by category of French text.**

Formal		Informal		Programming	
Char.	Freq. (%)	Char.	Freq. (%)	Char.	Freq. (%)
é	1.883	#	1.139	.	1.584
,	0.896	é	1.074	-	1.315
.	0.796	/	0.895	(	1.310
’	0.765	.	0.805	)	1.309
à	0.332	!	0.712	;	1.158
-	0.262	@	0.648	=	1.035
è	0.241	:	0.497	_	1.002
)	0.156	’	0.457	,	0.926
(	0.141	,	0.447	:	0.922
:	0.135	à	0.269	"	0.918
’	0.118	-	0.209	>	0.527
ê	0.098	"	0.185	/	0.459
/	0.078	è	0.155	<	0.445
!	0.075	’	0.129	{	0.444
;	0.058	ê	0.099	}	0.443
"	0.047	_	0.079	’	0.292
»	0.041	;	0.075	[	0.186
ç	0.041	&	0.068	]	0.186
<	0.041	)	0.063	%	0.150
?	0.040	<<	0.059	+	0.144

Interesting differences are visible. For instance, the mostly Internet-related characters # and @ appear in the table only for the Informal class. The common accented letters é, à, è, and ê are less frequent in Informal text than in the Formal corpora, although retaining the same relative order. Interestingly, / is present in all three columns, because of its wide range of uses.

iteratively “on the fly” as they become relevant. The idea is as follows: we start with the easy-to-solve linear part of the objective and ignore any quadratic terms at first. Iteratively, we generate the RLT1 relaxation of those quadratic terms  $a_{ijk} \cdot x_{ik}x_{jl}$  where at least one of the variables  $x_{ik}$  and  $x_{jl}$  is set to 1 in the previous optimal solution and where  $a_{ijk}$  has a substantial contribution to the objective value; in particular, we do not generate any  $y_{ijk}$  where  $a_{ijk} = 0$ . We thereby take advantage of the sparsity of the quadratic objective function, which allows us to introduce only a few additional variables in every iteration. After enhancing our model with these variables, we reoptimize until the addition of further terms does not significantly increase the objective value and the desired optimality gap is reached. This algorithm provides a hierarchy of lower bounds with every iteration producing a bound that is at least as good as the one from the previous iteration. For the problem instance that led to the final standardized layout, we could thus demonstrate that a very small gap (<2%) exists between the computed and the optimal solution after

only five days computation time. Note that, thanks to the sparsity, the formulations used in every iteration stay relatively small, enabling us to solve larger problem instances with less time and memory than the traditional complete RLT1 relaxation.

### Introducing Optimization Tools in the Standardization Process

The optimization approach described permits a principled approach to solving the keyboard layout problem. However, we quickly learned that a one-shot approach to optimization is not actionable in a complex, multi-stakeholder design project. The problem definition and expectations from stakeholders were ill-defined in the beginning and constantly evolving: definitions, parameters, and objectives changed, and decisions often hinged on subjective opinions, public feedback, or cultural norms, making them hard to express mathematically. We therefore ended up developing several approaches that helped integrate computational methods into the operational mode of the standardization committee.

When we first joined the project, the committee was debating each


character in hand-crafted layouts designed by individual members, with rationales such as

- “ê is **frequent**, so I gave it direct access because it’s **faster**.”
- “The guillemets (« ») are **important**, so they should be **easy to find**.”
- “@ **looks like** a, so I’ve put them **close together**.”
- “We should leave ç and ù where they are; otherwise, they will be **hard to find**.”


Many of these rationales were based on intuition, even when the objective measurement (of frequency, speed, and so on) was possible. Our first challenge in defining the optimization problem was to turn these rationales into well-defined quantified objectives. These hand-crafted proposals were typically good in one sense (for example, aiming for speed) but compromising other objectives. They often generated ideas following a greedy approach: starting with what seemed important and then having to make do with the remaining free slots and characters. The outcome of such a process depends greatly on the choice order, and on the subjective weights given to each rationale, which could vary hugely between characters and stakeholders.

Our first task was to explain how a combinatorial approach can assist with such complex, multi-criterion problems. In contrast to *ad-hoc* designs, formulating the problem in quantifiable objective metrics allows algorithms to consider all objectives at once and explore all possible solutions. It also enables stakeholders to assign understandable weights to the task’s many parameters, permitting exact control of their priorities. Also, the objective metrics can be evaluated separately for assessing effects of manual changes; room is left for design decisions based on subjective criteria that cannot be formalized.

We built an evaluation tool that replicated the objective criteria calculations used by the optimizer and used it to quickly compare competing layouts for different objectives. This allowed us to illustrate how easily character-by-character layout design can lead to suboptimal results. For example, evaluating one of the handmade layouts



**The challenge for us was to translate goals such as “facilitate typing and learning” into quantifiable objective functions.**

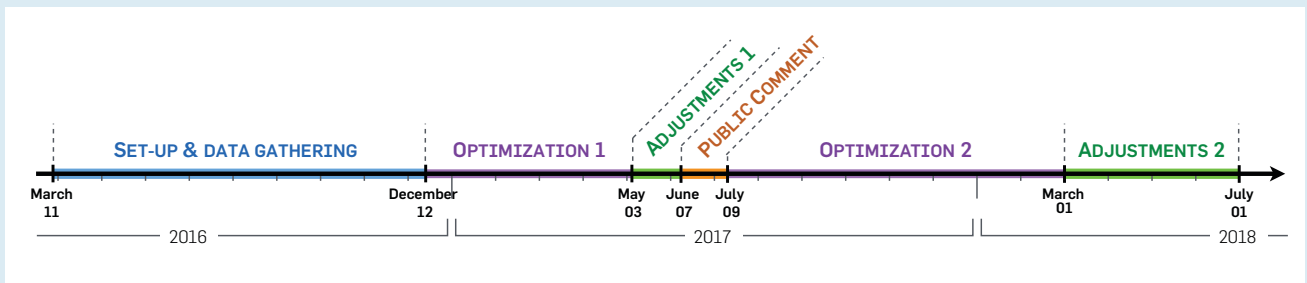


with the objective functions described above, we found that typing special characters was 47% slower, 48% less ergonomic, and 17% less similar to the traditional AZERTY than was our final optimized solution, which formed the basis for the new layout.

Over the course of the project, there were two cycles of *optimizations* and *adjustments*, separated by public consultation (see Figure 3). Before that, over nine months, we defined and iterated the optimization model with the committee, formulating objective functions that matched members’ intuitions and expectations (Table 1) and collected the text corpora. As we collected the input data, more subjective choices, such as character similarity, were discussed with the committee members and continuously adjusted over the course of the project. The first *optimization* phase entailed a five-month back-and-forth process between optimization and committee discussions. The optimizer computed solutions to numerous instances of the design problem, which we presented to the committee, explaining how inputs and constraints affected aspects of the designs. Members then suggested particular parameter settings or adding or removing constraints (for example, keeping capital and lowercase letters on the same key, changing the character sets, or weighting specific text corpora differently). We then optimized new layouts for these new parameters. After several such iterations, the committee agreed on the layout and parameter set it deemed best with regard to the optimization objectives.

Then, in the first *adjustments* phase, we used the optimizer to evaluate manual changes proposed by the experts. It was argued that these adjustments capture exceptions to the objectives, such as individuals’ expectations and preferences, cultural norms, or character-specific political decisions that frequently changed with every iteration and could not be formally modeled. For example, the traditional position for the underscore was preferred for some solutions, thanks in part to nomenclature: it is colloquially called the “8’s dash” (*tiret du 8*), for its location on the 8 key in AZERTY. The aforementioned evaluation tool helped us assess the consequences of these character moves or swaps on the four objective criteria.

**Figure 3. Project timeline: computational methods were involved in all phases but the public comment and were governed by interactions with stakeholders and various computational methods developed by the researchers.**



The committee hence could make better-informed decisions about trade-offs between adjustments. This led to the first release candidate.

In June 2017, this layout was presented to the public, which had 1 month, per AFNOR’s standard procedure, to respond to the proposed standard and offer comments and suggestions. An unprecedented number of responses (over 3,700) were submitted, including numerous suggestions. Feedback was strongly divided on some matters, such as how strongly computer-programming-related characters should be favored, or where accented characters should be placed. The committee compiled the feedback into themes and tried to identify consensual topics. In some cases, there were opposing sides with no clear majority. For example, some people insisted that all pre-marked characters (for example, é à ç) be removed from the layout to make other characters more accessible, because the former could be entered using combining accents, whereas others argued for having even more pre-accentuated letters accessible directly. Consensus itself could also be difficult to assess: a subset of people argued that digits should be accessible without the Shift modifier, but it was not clear whether all of the remaining commenters were positive, neutral, or even gave any thought to keeping them “shifted.” In such cases, the committee referred to the Ministry of Culture’s stated objectives as well as to the experts’ opinions on the available options: digits in our text corpora are much less frequent than some of the most used accented characters, and the change from the traditional AZERTY was deemed too large.

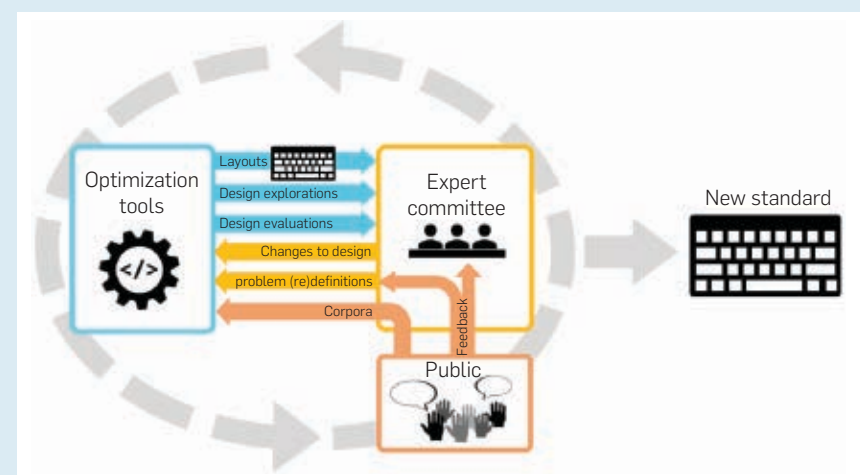
Consensual trends in the comments directly led to updates of the optimization model, its inputs, or parameters: characters were added or removed from the initial set; some associations were added to the Intuitiveness criterion and the weights of the criteria and corpora were updated. Hard constraints were added to the optimizer, such as having opening and closing character pairs (for example, [ ] { } “ ” « ») placed on consecutive keys on the same row and with the same modifiers. Finally, the positions of @ and # were fixed to more accessible slots already used in alternative AZERTY layouts.

The second cycle then began, consisting of a (seven-month) *optimization* and (four-month) *adjustment* phase, similar to the ones described above.

Figure 4 summarizes our approach to integrate computational methods into the standardization of the French

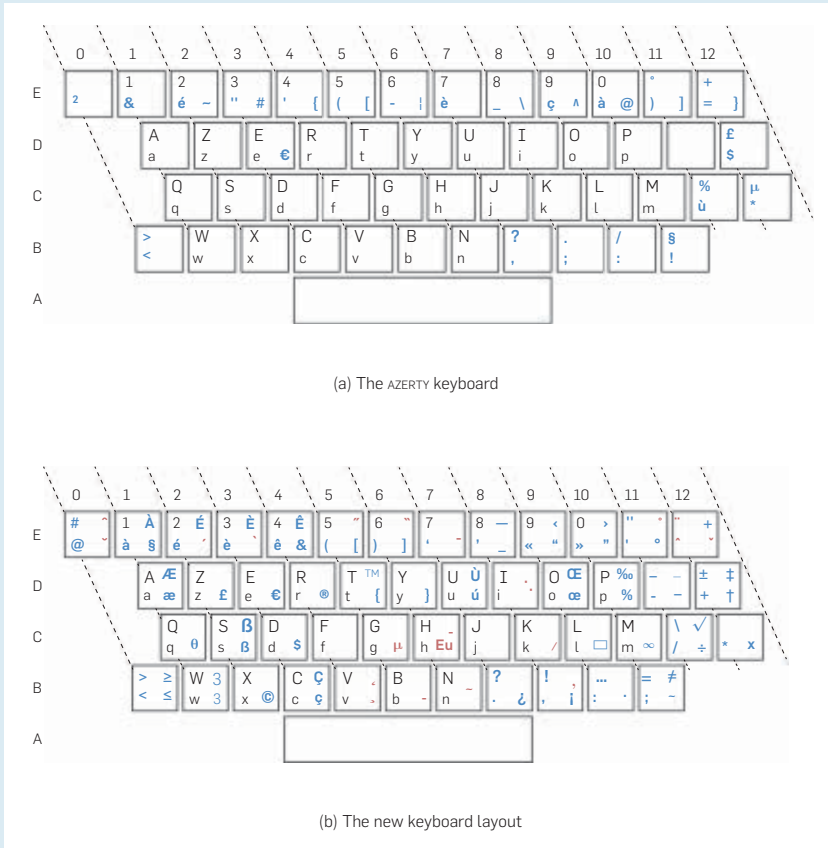
keyboard and diagrams the interactions we developed. Our optimization tools proposed solutions and could be used to evaluate suggestions, which enabled an efficient exploration of the very large design space. On the other side, the experts steered the definition of objectives, set weights, and adjusted the input data. They used the optimized designs to explore the solution space and tweaked the computed layouts to consider tacit criteria too, such as political objectives and cultural norms. The evaluation tool could be used to study the consequences of conflicting views, for example, by quickly checking what happens to objective scores when a character is moved. Simultaneously, both sides were informed by comments from the public, whose expectations and wishes led the experts to question their assumptions and criteria and were directly implemented as

**Figure 4. Diagram of our participatory optimization process wherein experts on a standardization committee define objectives and inputs to an optimizer, which, in turn, supplies concrete layouts, accompanied with feedback on quality (performance and intuitiveness, among others). The process was informed by feedback from the public.**





**Figure 5. Comparison of the AZERTY and the new standardized layout. The characters included in the design problem are in boldface and color. Marked in red are dead keys, which require pressing a subsequent key before a symbol is produced (diacritical marks and mode keys for accessing non-French-language Latin characters, Greek letters, and currency symbols).**



changes in the weight and constraint definitions within the optimization model.

In summary, customized tools applying an established optimization approach allowed fast iteration and explainable results, and provided monitoring tools that enabled stakeholders to test and assess the effects of their ideas for every measurable objective goal, yielding transparent results. We arrived at the final layout by combining objective and subjective criteria weighted and refined through several iterations with computational tools. This involved hard facts whenever possible and factoring in numerous opinions not only from diverse experts but also from the public, the primary target of the new standard.

### The New French Keyboard Standard

The outcome, shown in Figure 5b, makes it easier to type French and

enables accessing a larger set of characters. Despite the problem’s computational complexity, we were able to propose a solution for which we could computationally verify that it is within 1.98% of the best achievable design with regard to the overall objective function and the final choice of parameters presented in Table 1. This means that it is either optimal or, if suboptimal, at most 1.98% worse than an unknown optimal design.<sup>13</sup> This solution was taken as a design basis, to which the committee added 24 further, rarer characters. Manual changes were made to accommodate these and locally optimize the layout’s intuitiveness. All decisions were informed by our evaluation tool, allowing the committee to finely control the consequences of each manual change to the initial four objectives.

The new layout enables direct input of more than 190 special characters, a significant increase from the 47 of the current

AZERTY.<sup>c</sup> It allows accessing all characters used in French without relying on software-side corrections. Frequently used French characters are accessible without any modifier (é, à, «, », and so on), or intuitively positioned where users can expect them (for example, œ on the o key). All accented capital letters (À and É, among others) can be entered directly or using a dead key. The main layout offers almost 60 characters not available in AZERTY for entering symbols used in math, linguistics, economics, programming, and other fields. Some programming characters, which often have alternative uses, were given more prominent slots; for instance, / became accessible without modifiers and \ is on the same key but in a shifted slot. According to the metrics described above, the performance and ergonomics of typing the special characters already present in traditional AZERTY are improved by 18.4% and 8.4%, respectively, even though the new layout had to accommodate 60 additional characters.

The keyboard offers three additional layers accessed via special mode keys. These are dedicated to European characters not used in French (via the Eu key from Alt+H in Figure 5b), currency symbols (via ₤ with Alt+F), and Greek letters (via Alt+G’s μ), more than 80 additional characters in all. Their placement was beyond the scope of the optimization process, being near-nonexistent in our text corpora.

Its many changes notwithstanding, the layout maintains similarity to the traditional AZERTY, making the transition for users simple. Of the 45 special characters previously available, 8 retained their original location and 12 moved by less than three keys. In particular, frequently used characters were kept near their original position. For instance, the most common special character (é) is not in the fastest spot to access on average (B07 in our study) but stayed at E02 for similarity although maintaining good performance. Many punctuation characters (slots B7–B10) were moved slightly by the optimizer to better reflect character and character-pair frequencies (see Table 2) although remaining in the expected area of the keyboard. Comparing the final design to AZER-

<sup>c</sup> Not including accented characters that can be created using dead keys, such as  $\hat{a} + \hat{I} = \hat{a}$ .

TY based on our objective functions, we can see that all larger moves of characters had a clear justification, be it better performance, ergonomics, discoverability, or consistency. Most noteworthy was bringing paired characters such as parentheses and brackets closer together, a direct result of the public consultation.

Finally, substantial effort was devoted to forming semantic regions for characters, such as mathematical characters (C11–D12 and B12), common currency symbols (C02–D03), or quotation marks (E07–E11). Many of these groupings emerged during the optimization process, thanks to the Intuitiveness objective. Others resulted from manual changes when the committee decided to prioritize semantic grouping over performance or ergonomics (for example, following a calculator metaphor for mathematical characters). The Intuitiveness score improved more than fourfold (434.4%) relative to the traditional AZERTY.

**Communication and adoption.** We cannot predict the success of the new standard, nor how quickly users will adapt it. Being voluntary, its publication does not bind users nor manufacturers. We can, however, report first indicators of interest, as well as the French Ministry of Culture’s plans to promote the new layout.

At least two manufacturers started producing physical keyboards engraved according to the new standard, of which already one was marketed by the end of 2019. We were also informed that Microsoft will integrate an official driver to Windows 10. Importantly, as an attempt to promote the use of the new layout, the French Ministry of Culture reported that they will replace the entire “fleet” of its employees’ keyboards. We also received numerous emails from individuals motivated to write their own keyboard drivers and key-stickers, so they and others could use the layout before it is effectively commercialized. Only few months after the release of the standard, several drivers were available for Mac OSX, Windows 10, and Linux; some of them listed on our webpage.<sup>d</sup> These measures indicate the will and potential for nationwide adoption.

To inform users and encourage pub-

lic acceptance, we published an interactive visualization of the keyboard online,<sup>d</sup> in which people can explore to discover the new layout and learn the reasoning behind it. It received more than 74,800 page views in the week following the official release event on April 2, 2019, and counted more than 122,000 views 5 months after the standard was published. For people interested in finer details, we also published an open-access document in French and English explaining the essence of our method in layman’s terms.<sup>8,19</sup> This details the impact of the various corpora and weights involved in the calculations and in the committee’s later deliberations.

**Learnings and Outlook**

The design of keyboards is a matter of economic, societal, and even medical interest. However, as most complex artifacts involving software do, they evolve by stacking layers on layers. Most keyboard layouts were designed decades ago or more. To respond to chang-

ing uses of language, from programming to social media, they have evolved incrementally via adding characters to unused keyboard slots. The absence of appropriate layout standards negatively affects the preservation and evolution of these languages. Indeed, it is startling that some of the world’s most spoken languages<sup>21</sup> lack any government-approved keyboard standard: Punjabi (10<sup>th</sup>), Telegu (15<sup>th</sup>), and Marathi (19<sup>th</sup>).

Similarly, virtual (software) keyboards mostly follow agreed-on standards for alphanumeric characters, but special characters can be company-specific and vary greatly. Computational design methods could play a role in helping regulators improve quality and respond more swiftly to changes in computing and language, even “shaking up” a design if needed. The optimization methods and tools proposed here can be applied to other languages and input methods (for example, touchscreens) with adaptations to the input data and corresponding weights.

For keyboards and beyond, we be-

**Table 3. Opportunities for improving the use of computational methods in large-scale design projects, identified on the basis of our experience in using combinatorial optimization for designing the French keyboard standard.**

**Facilitating participatory optimization**

To support multistakeholder design projects, computational methods should be interactive, iterative, and participatory. Therefore we need tools that allow:

(1) **Fast (re)definition of the problem:**

In an iterative design process, the problem definition is constantly evolving. To speed up computation in cases of only slight changes in definition or instances, standard solvers should find a way to reuse information about previously explored solutions, as with the pruning decisions in a branch-and-bound tree, and adapt them to the modified constraints and objectives.

(2) **Online exploration of the design space:**

Manual exploration is essential for stakeholders’ understanding of the design problem and speculation such as “what if we group all math characters on the right side of the keyboard?” General-purpose solvers lack interfaces for manually exploring the design space.<sup>9,15</sup> A two-way interface is needed that lets stakeholders change solutions or propose new ones and enables the optimization process to communicate the outcome from the assessment in human-readable format.

(3) **Learning and visualizing hidden “subjective” criteria:**

Our stakeholders made manual adjustments to a proposed solution, applying tacit criteria such as assumptions about users’ habits, cultural specificities, subjective preferences, and political agendas. Optimizers should offer interfaces for making such local changes. From the interactions, a “subjective function” could be learned that can be shown and used as an additional objective in optimization of future solutions.

(4) **Justifications for design choices:**

When presented with a solution, the committee and the public often asked questions of the form “why is this character placed here?” and wanted to understand how a change in the objective weight or parameters would affect the optimal solution. Developing effective visualizations that show how changes to optimization parameters impact the design and *vice versa* could aid users in navigating the design space and make the optimization more predictable.

<sup>d</sup> See <http://norme-AZERTY.fr/>

lieve that much of the potential of computational methods remains unexploited. The power of algorithms lies in their problem-solving capability. They can explore design spaces and obtain suggestions that would be hard to find by intuition or trial-and-error. This element is often missing from present-day mainstream interaction design, which leaves the generation of new designs to humans.

However, the case of the French keyboard has revealed important challenges in integrating computational methods into large-scale multi-stakeholder design projects. Starting from a well-defined optimization problem, our approach evolved toward something one could call *participatory optimization*. This is inspired by participatory design, which originated with labor unions and was developed as a co-design method aimed at democratic inclusion of stakeholders.<sup>23</sup> Equal representation and resolving conflicts were two key aims. For such optimization, the stakeholders must be brought together at a level where they can inform and influence each other interactively and iteratively, engaging directly with the optimizer and model to arrive at a good solution collaboratively.

There is growing interest in optimization research employing methods that actively include the user in the process. However, the notion of participatory optimization goes beyond previous efforts to simply open up the search- and model-building process for input by the end-user.<sup>18</sup> It focuses particularly on including stakeholders at every step in the process, for which state-of-the-art optimization methods provide limited support. The case of the French keyboard reveals 4 avenues for future work as especially important to address for enabling active participation of stakeholders and optimizer in an iterative human-centered design process supported by computational methods (see Table 3).

We envision such demonstrations as ours encouraging establishment of new, human-centered objectives in algorithm research. Considering interactive and participatory properties of algorithms also opens new questions and paths to new, societally important uses. How well can we stop, refine, and resume an algorithm? Can we define

task instances in different ways and leave some variables open? Can we visualize the search landscape meaningfully, or learn “subjective functions” from interactions? Can we use fast approximations in lieu of full-fledged solvers in interactive design sessions? We believe that when designed from a participatory perspective, algorithms could more directly support not only problem-solving but also considering multiple perspectives, making refinements, and learning about a problem.

The code and data presented in this article are documented and open-sourced,<sup>e</sup> alongside instructions for optimizing a layout for any language. **□**

e See <http://norme-AZERTY.fr>.

References

1. Bertsimas, D., Tsitsiklis, J. *Introduction to Linear Optimization*, 1<sup>st</sup> edn. Athena Scientific, Belmont, MA, USA, 1997. ISBN 1886529191.
2. Burkard, R.E., Offermann, D.M.J. Entwurf von Schreibmaschinentastaturen mittels quadratischer Zuordnungsprobleme. *Zeitschrift für Opt. Res.* 21, 4 (1977), B121–B132.
3. Burkard, R.E., Karisch, S.E., Rendl, F. Qaplib - a quadratic assignment problem library. *J. Global Optim.* 10, 4 (June 1997), 391–403. ISSN 0925-5001. doi: 10.1023/A:1008293323270. <http://dx.doi.org/10.1023/A:1008293323270>.
4. DGLFLF. *Rapport au Parlement sur l'emploi de la langue française*. Government Report, 2015. <http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/La-DGLFLF/Nos-priorites/Rapport-au-Parlement-sur-l-emploi-de-la-langue-francaise-2015>. From the Délégation générale à la langue française et aux langues de France of the Ministère de la Culture et de la Communication. In French.
5. DGLFLF. *Vers une norme française pour les claviers informatiques*. Government Publication, 2016. <http://www.culture.gouv.fr/Thematiques/Langue-francaise-et-langues-de-France/Politiques-de-la-langue/Langues-et-numerique/Les-technologies-de-la-langue-et-la-normalisation/Vers-une-norme-francaise-pour-les-claviers-informatiques>. From the Délégation générale à la langue française et aux langues de France of the Ministère de la Culture et de la Communication. In French.
6. Feit, A.M. Assignment problems for optimizing text input. PhD thesis, Aalto University, 2018. <http://urn.fi/URN:ISBN:978-952-60-8016-1>.
7. Feit, A.M., Weir, D., Oulasvirta, A. How we type: Movement strategies and performance in everyday typing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16, 2016). ACM, New York, NY, USA, 4262–4273. ISBN 978-1-4503-3362-7. doi: 10.1145/2858036.2858233. <http://doi.acm.org/10.1145/2858036.2858233>.
8. Feit, A.M., Nancel, M., Weir, D., Bailly, G., John, M., Karrenbauer, A., Oulasvirta, A. *Elaboration de la disposition AZERTY modernisée*. Technical report, Inria, Lille, France 2018. <https://hal.inria.fr/hal-01826476>.
9. Fisher, M.L., Rosenwein, M.B. An interactive optimization system for bulk-cargo ship scheduling. *Nav. Res. Logist.* 36, 1 (1989), 27–42.
10. Frieze, A.M., Yadegar, J. On the quadratic assignment problem. *Discrete Appl. Math.* 5, 1 (1983), 89–98. ISSN 0166-218X. doi: [http://dx.doi.org/10.1016/0166-218X\(83\)90018-5](http://dx.doi.org/10.1016/0166-218X(83)90018-5). <http://www.sciencedirect.com/science/article/pii/0166218X83900185>.
11. Gajos, K., Weld, D.S. Preference elicitation for interface optimization. In *Proceedings of the 18<sup>th</sup> annual ACM symposium on User interface software and technology* (2005), ACM, New York, NY, USA, 173–182.
12. ISO/IEC 9995-1. *ISO/IEC 9995-1:2009 Information technology – Keyboard layouts for text and office systems – Part 1: General principles governing keyboard layouts*. Standard, International Organization

- for Standardization, Geneva, CH, October 2009.
13. John, M., Karrenbauer, A. Dynamic sparsification for quadratic assignment problems. In *Mathematical Optimization Theory and Operations Research*. M. Khachay, Y. Kochetov, P. Pardalos, eds. Springer International Publishing, Cham, 2019, 232–246. ISBN 978-3-030-22629-9.
14. Jokinen, J.P.P., Sarcar, S., Oulasvirta, A., Silpasuwanchai, C., Wang, Z., Ren, X. Modelling learning of new keyboard layouts. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17, 2017), ACM Press, New York, New York, USA, 4203–4215. ISBN 9781450346559. doi: 10.1145/3025453.3025580. <http://dl.acm.org/citation.cfm?doi=3025453.3025580>.
15. Kapoor, A., Lee, B., Tan, D., Horvitz, E. Interactive optimization for steering machine classification. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10, 2010), ACM, New York, NY, USA, 1343–1352. ISBN 978-1-60558-929-9. doi: 10.1145/1753326.1753529. <http://doi.acm.org/10.1145/1753326.1753529>.
16. Karrenbauer, A., Oulasvirta, A. Improvements to keyboard optimization with integer programming. In *Proceedings of the 27<sup>th</sup> Annual ACM Symposium on User Interface Software and Technology* (UIST '14, 2014), ACM, New York, NY, USA, 621–626. ISBN 978-1-4503-3069-5. doi: 10.1145/2642918.2647382. <http://doi.acm.org/10.1145/2642918.2647382>.
17. Lee, P.U.-J., Zhai, S. Top-down learning strategies: can they facilitate stylus keyboard learning? *Int. J. Human-Computer Stud.* 60, 5-6 (May 2004), 585–598. ISSN 1071-5819. doi: 10.1016/J.IJHCS.2003.10.009. <http://www.sciencedirect.com/science/article/pii/S1071581903001794>.
18. Meignan, D., Knust, S., Frayret, J.-M., Pesant, G., Gaud, N. A review and taxonomy of interactive optimization methods in operations research. *ACM Trans. Interact. Intell. Syst.*, 5(3):17:1–17:43, September 2015. ISSN 2160-6455. doi: 10.1145/2808234. URL <http://doi.acm.org/10.1145/2808234>.
19. NF Z 71-300. *User interfaces - French keyboard layouts for office*. Standard, AFNOR, 03, 2019a.
20. NF Z 71-300. *Interfaces utilisateurs - Dispositions de clavier bureautique français*. Standard, AFNOR, 03, 2019b.
21. Parkvall, M. Vårdens 100 största språk 2007. *Nationalencyklopedin*, 2007. <http://www.ne.se/>.
22. Queyranne, M. Performance ratio of polynomial heuristics for triangle inequality quadratic assignment problems. *Oper. Res. Lett.* 4, 5 (1986), 231–234. ISSN 0167-6377. doi: 10.1016/0167-6377(86)90007-6. <http://www.sciencedirect.com/science/article/pii/0167637786900076>.
23. Simonsen, J., Robertson, T. *Routledge International Handbook of Participatory Design*. Routledge, New York, NY, USA, 2012.
24. Yassi, A. Repetitive strain injuries. *The Lancet* 349, 9056 (March 1997), 943–947. doi: 10.1016/S0140-6736(96)07221-2. <http://www.sciencedirect.com/science/article/pii/S0140673696072212>. Manuscript submitted to ACM

**Anna Maria Feit** ([feit@cs.uni-saarland.de](mailto:feit@cs.uni-saarland.de)) is a professor at Saarland University, Germany. This work was done while a researcher at Aalto University and ETH Zurich, Switzerland.

**Mathieu Nancel** is a research scientist in the Loki research group at Inria Lille–Nord Europe; Lille, France.

**Maximilian John** is a researcher at Max Planck Institute for Informatics, Saarbrücken, Germany.

**Andreas Karrenbauer** is a senior researcher at Max Planck Institute for Informatics, Saarbrücken, Germany.

**Daryl Weir** is a researcher at Aalto University, Espoo, Finland.

**Antti Oulasvirta** is an associate professor at Aalto University, Espoo, Finland.

© 2021 ACM 0001-0782/21/2 \$15.00



Watch the authors discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/azerty-ameliore>



---

## How design and operation of modern cloud-scale systems conflict with GDPR.

---

BY SUPREETH SHASTRI, MELISSA WASSERMAN,  
AND VIJAY CHIDAMBARAM

---

# GDPR Anti- Patterns

THE GENERAL DATA Protection Regulation (GDPR)<sup>26</sup> is a European privacy law introduced to offer new rights and protections to people concerning their personal data. While at-scale monetization of personal data has existed since the dot-com era, a systemic disregard for privacy and protection of personal data is a recent

phenomenon. For example, in 2017, we learned about Equifax's negligence<sup>17</sup> in following the security protocols, which exposed the financial records of 145 million people; Yahoo!'s delayed confession<sup>21</sup> that three years ago, a theft had exposed all three billion of its user records; Facebook's admission<sup>33</sup> that their APIs allowed illegal harvesting of user data, which in turn influenced the U.S. and U.K. democratic processes.

Thus, GDPR was enacted to prevent a widespread and systemic abuse of personal data. At its core, GDPR declares the privacy and protection of

personal data as a fundamental right. Accordingly, it grants new rights to people, and assigns companies that collect their personal data, new responsibilities. Any company dealing with the personal data of European people is legally bound to comply with all the regulations of GDPR, or risk facing hefty financial penalties. For example, in January 2019, Google was fined<sup>6</sup> €50M for lacking a customer's consent in personalizing advertisements across their different services.

In this work, we investigate the challenges that modern cloud-scale systems face in complying with GDPR.

Specifically, we focus on the design principles and operational practices of these systems that conflict with the requirements of GDPR. To capture this tussle, we introduce the notion of GDPR *anti-patterns*. In contrast to outright bad behavior, say storing customer passwords in plaintext, GDPR anti-patterns are those practices that serve their originally intended purpose well but violate the norms of GDPR. For example, given the commercial value of personal data, modern systems have naturally evolved to store them without a clear timeline for deletion, and to reuse them across various applications. While these practices help the systems generate more revenue and thereby value, they violate the storage and purpose limitations of GDPR.

Building on our work analyzing GDPR from a systems perspective,<sup>30–32</sup> we identify six GDPR anti-patterns that are widely present in the real world. These include storing personal data without a timeline for deletion; reusing personal data indiscriminately; creating black markets for personal data; risk-agnostic data processing; hiding data breaches; and making unexplainable decisions. These anti-patterns highlight how the traditional system design goals of optimizing for performance, cost, and reliability sit at odds with GDPR’s goal of data protection by design and by default. While eliminating these anti-patterns is not enough to achieve overall compliance under GDPR, ignoring these will definitely violate its intents.

We structure the rest of this article as follows: First, we provide a brief primer on GDPR, then describe the six GDPR anti-patterns, discussing how they came to be, reviewing the conflicting regulations, and chronicling their real-world implications. Finally, we ruminate on the challenges and opportunities for system designers as societies embrace data protection regulations.

**GDPR**

On May 25, 2018, the European Parliament rolled out the General Data Protection Regulation.<sup>26</sup> In contrast with targeted privacy regulations like HIPAA and FERPA in the United States, GDPR takes a comprehensive view by defining *personal data* to be any information relating to an identifiable natural person. GDPR defines three entities that interact with personal data: *data subject*, the person whose personal data is collected; *data controller*, the entity that collects and uses personal data; and, *data processor*, the entity that processes personal data on behalf of a data controller. Then, GDPR designates supervisory authorities (one per EU country) to oversee that the rights and responsibilities of GDPR are complied with.

The accompanying figure represents how GDPR entities interact with each other in collecting, storing, processing, securing, and sharing personal data. Consider the music streaming company Spotify collecting its customers’ listening history, and then using Google cloud’s services to determine

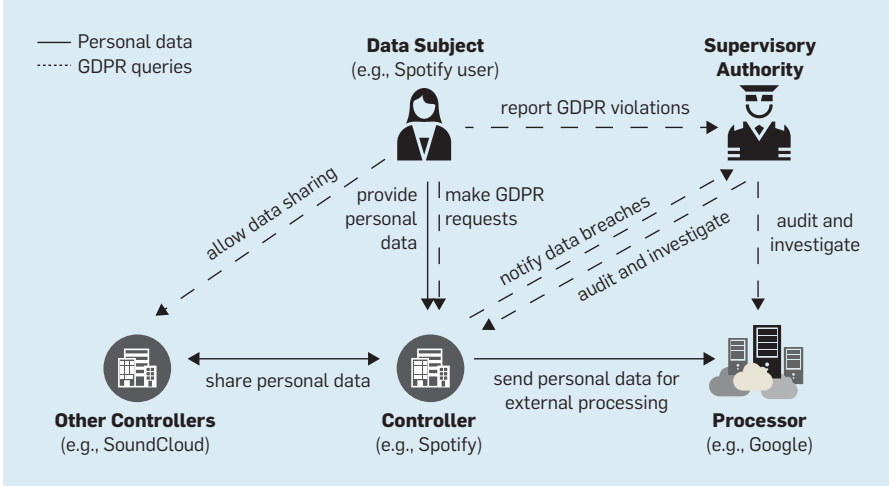
new recommendations for customers. In this scenario, Spotify is the data controller and Google Cloud is the data processor. Spotify could also engage with other data controllers, say SoundCloud, to gather additional personal data of their customers.

To ensure privacy and protection of personal data in such ecosystems, GDPR grants new rights to customers and assigns responsibilities to controllers and processors. Now, any person can request a controller to grant access to all their personal data, to rectify errors, to request deletion, to object to their data being used for specific purposes, to port their data to third parties and so on. On the other hand, the controller is required to obtain people’s consent before using their personal data, to notify them of data breaches within 72 hours of finding out, to design systems that are secure by design and by default, and to maintain records of activities performed on personal data. For controllers failing to comply with these rights and responsibilities, GDPR regulators could levy penalties of up to €20M or 4% of their annual global revenue, whichever is higher.

*Structure.* GDPR is organized as 99 articles that describe its legal requirements, and 173 recitals that provide additional context and clarifications to these articles. The first 11 articles layout the principles of data privacy; articles 12–23 establish the rights of the people; then articles 24–50 mandate the responsibilities of the data controllers and processors; the next 26 articles describe the role and tasks of supervisory authorities; and the remainder of the articles cover liabilities, penalties and specific situations. We expand on the relevant articles later.

*Impact.* Compliance with GDPR has been a challenge for many companies that collect personal data. A number of companies like Klout and Unroll.me terminated their services in Europe to avoid the hassles of compliance. Few other businesses made temporary modifications. For example, media site *USA Today* turned off all advertisements, whereas the *New York Times* stopped serving personalized ads. While most organizations are working toward compliance, Gartner reports<sup>13</sup>

**Flow of personal data and GDPR queries between the four GDPR entities: data subjects, data controllers, data processors, and regulators.**





that less than 50% of the companies affected by GDPR were compliant by the end of 2018. This challenge is further exacerbated by the performance impact that GDPR compliance imposes on current systems.<sup>30</sup>

In contrast, people have been enthusiastically exercising their newfound rights. In fact, the EU data protection board reports<sup>12</sup> having received 144,376 complaints from individuals and organizations in the first year of GDPR. Surprisingly, even the companies have been forthcoming in reporting their security failures and data breaches, with 89,271 breach notifications sent to regulators in the same 12-month period. In 2019, several companies have been levied hefty penalties for GDPR violations: €50 million for Google,<sup>6</sup> £99M for Marriott International,<sup>25</sup> and £183M for British Airways.<sup>24</sup>

### GDPR Anti-Patterns

The notion of anti-patterns was first introduced<sup>19</sup> by Andrew Koenig to characterize patterns of software design and behavior that superficially look like good solutions but end up being counterproductive in reality. An example of this is performing premature optimizations in software systems. Extending this notion, we define the term GDPR anti-patterns to refer to system designs and operational practices, which are effective in their own context but violate the rights and regulations of GDPR. Naturally, our definition does not include design choices that are bad in their own right, say storing customer passwords in plaintext, though they also violate GDPR norms. In this section, we catalog six GDPR anti-patterns, detailing how they came to be, which regulations they violate, and their implications in the real-world.

*Genesis.* GDPR anti-patterns presented here have evolved from the practices and design considerations of the post dot-com era (circa 2000). These modern cloud-scale systems could be characterized by their quest for unprecedented scalability, reliability, and affordability. For example, Google operates eight global-scale applications at 99.99% uptime with each of them supporting more than one billion users. Similarly, Amazon's cloud computing infrastructure provides on-demand access to inexpensive computing to over 1 million users in 190 countries, all the while guaranteeing four nines of availability. This unrelenting focus on performance, cost-efficiency, reliability, and scalability has resulted in relegating security and privacy to a backseat.

While our GDPR analysis recognizes six anti-patterns, this list is not comprehensive. There are many other



unsavory practices that would not stand the regulator scrutiny. For example, the design and operation of consent-free behavioral tracking.<sup>22</sup> Our goal here is to highlight how some of the design principles, architectural components, and operational practices of the modern cloud-scale systems conflict with the rights and responsibilities laid out in GDPR. We present six such anti-patterns and summarize them in the accompanying table.

**Storing data without a clear timeline for deletion.** Computing systems have always relied on insights derived from data. However, in recent years, this dependence is reaching new heights with a widespread adoption of machine learning and big data analytics in system design. Data has been compared to oil, electricity, gold, and even bacon.<sup>1</sup> Naturally, technology companies evolved to not only collect personal data aggressively but also to preserve them forever. However, GDPR mandates that no data lives without a clear timeline for deletion.

ARTICLE 17: RIGHT TO BE FORGOTTEN. *“(1) The data subject shall have the right to obtain from the controller the erasure of personal data without undue delay ...”*

ARTICLE 13: INFORMATION TO BE PROVIDED WHERE PERSONAL

DATA ARE COLLECTED FROM THE DATA SUBJECT. *“(2)(a) ... the controller shall provide the period for which the personal data will be stored, or the criteria used to determine that period;”*

ARTICLE 5(1)(E): STORAGE LIMITATION. *“kept... for no longer than is necessary for the purposes for which the personal data are processed ...”*

GDPR grants data subjects an unconditional right, via article 17, to request their personal data be removed from the system within a reasonable time. In conjunction with this, articles 5 and 13 lay out additional responsibilities for the data controller: at the point of collection, users should be informed the time period for which their personal data would be stored, and if the personal data is no longer necessary for the purpose for which it was collected, then it should be deleted. These simply mean that all personal data should have a time-to-live (TTL) that data subjects are aware of, and that controllers honor. However, the law makes exceptions for archiving data in the public interest, or for scientific or historical research purposes.

*Deletion in the real world.* While conceptually clear, a timely and guaranteed removal of data is challenging in

practice. For example, Google cloud describes the deletion of customer data as an iterative process<sup>8</sup> that could take up to 180 days to fully complete. This is because, for performance, reliability, and scalability reasons, parts of data get replicated in various storage subsystems like memory, cache, disks, tapes, and network storage; multiple copies of data are saved in redundant backups and geographically distributed datacenters. Such practices not only delay the timeliness of deletions but also make it harder to offer guarantees.

**Reusing data indiscriminately.** While designing software systems, a purpose is typically associated with programs and models, whereas data is viewed as a helper resource that serves these high-level entities in accomplishing their goals. This portrayal of data as an inert entity allows it to be used freely and fungibly across various systems. For example, this has enabled organizations like Google and Facebook to collect user data once and use it to personalize their experiences across several services. However, GDPR regulations prohibit this practice.

ARTICLE 5(1)(B): PURPOSE LIMITATION. *“Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes ...”*

ARTICLE 6: LAWFULNESS OF PROCESSING. *“(1)(a) Processing shall be lawful only if ... the data subject has given consent to the processing of his or her personal data for one or more specific purposes.”*

ARTICLE 21: RIGHT TO OBJECT. *“(1) The data subject shall have the right to object ... at any time to processing of personal data concerning him or her ...”*

The first two articles establish that personal data could only be collected for specific purposes and not be used for anything else. Then, article 21 grants users a right to object, at any time, to their personal data being

**GDPR anti-patterns, their real-world examples, and the GDPR articles that prohibit such behavior.**

Anti-Pattern	Real-World Examples	Governing GDPR Articles
Storing data without a clear timeline for deletion	Search engines before Right-to-be-forgotten (circa 2014)	5(1e). Storage limitation 17. Right to be forgotten
Reusing data indiscriminately	Facebook collecting phone numbers for 2FA and using them for ads and marketing	5(1b). Purpose limitation 6. Lawfulness of processing 21. Right to object
Creating black markets	Illegal data harvesting by programmatic ad exchanges	14. Information to be provided[...] 20. Right to data portability
Risk-agnostic data processing	Strava global heatmap that revealed classified military bases	35. Data protection impact assessment 36. Prior consultation
Hiding data breaches	Uber paying off hackers to hide their 2016 data breach	5. Principles relating to processing 33. Notification of personal data breach
Making unexplainable decisions	Using software like COMPASS in courts to predict recidivism	15. Right of access by the data subject 22. Automated individual decisionmaking

used for any purpose including marketing, scientific research, or historical archiving, or profiling. Together, these articles require each personal data (or groups of related data) to have their own blacklisted and whitelisted purposes that could be changed over time.

*Purpose in the real world.* The impact of the purpose requirement has been swift and consequential. For example, in January 2019, the French data protection commission<sup>6</sup> fined Google €50M for not having a legal basis for their ads' personalization. Specifically, the ruling said the user consent obtained by Google was not "specific" enough, and the personal data thus obtained should not have been used across 20 services.

**Walled gardens and black markets.** As we are in the early days of large-scale commoditization of personal data, the norms for acquiring, sharing, and reselling them are not yet well established. This has led to uncertainties for people and a tussle for control over data among controllers. People are concerned about vendor lock-ins, and about a lack of visibility once their data is shared or sold in secondary markets. Organizations have responded to this by setting up walled gardens and making secondary markets even more opaque. However, GDPR dismantles such practices.

ARTICLE 20: RIGHT TO DATA PORTABILITY. "(1) *The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller.* (2) *... the right to have the personal data transmitted directly from one controller to another.*"

ARTICLE 14: INFORMATION TO BE PROVIDED WHERE PERSONAL DATA HAVE NOT BEEN OBTAINED FROM THE DATA SUBJECT. "(1) (c) *the purposes of the processing ..., (e) the recipients ..., (2) (a) the period for which the personal data will be stored ..., (f) from which source the personal data originate ... (3) The controller shall provide the information at the latest within one month.*"

With article 20, people have a right to request for all the personal data that a controller has collected directly from them. Not only that, they could also ask the controller to directly transmit all such personal data to a different controller. While that tackles the vendor lock-ins, article 14 regulates the behavior in secondary markets. It requires that anyone indirectly procuring personal data must inform the data subjects, within a month, about how they acquired it, how long would they be stored, what purpose would they be used for, and who they intend to share it with. The *data trail* set up by this regulation should bring control and clarity back to the people.

*Data movement in the real world.* When GDPR went live, a large number of companies rolled out<sup>7</sup> data download tools for EU users. For example, Google Takeout lets users not only access all their personal data in their system but also port data directly to external services. However, the impact has been less savory for programmatic ad exchanges<sup>9</sup> in Europe, many of which had to shut down. This was primarily due to Google and Facebook restricting access to their platforms for those ad exchanges, which could not verify the legality of the personal data they possessed.

**Risk-agnostic data processing.** Modern technology companies face the challenge of creating and managing increasingly complex software systems in an environment that demands rapid innovation. This has led to a practice, especially in the Internet-era companies, of prioritizing speed over correctness; and to a belief that *unless you are breaking stuff, you are not moving fast enough*. However, GDPR explicitly restricts such approaches when dealing with personal data.

ARTICLE 35: DATA PROTECTION IMPACT ASSESSMENT. "(1) *Where processing, in particular using new technologies, ... is likely to result in a high risk to the rights and freedoms of natural persons, the controller shall, prior to the processing, carry out an assessment of the impact of the envisaged processing operations on the protection of personal data.*"

ARTICLE 36: PRIOR CONSULTATION. "(1) *The controller shall consult the supervisory authority prior to processing where ... that would result in a high risk in the absence of measures taken by the controller to mitigate the risk.*"

GDPR establishes, via articles 35 and 36, two levels of checks for introducing new technologies and for modifying existing systems, if they process large amounts of personal data. The first level is internal to the controller, where an impact assessment must analyze the nature and scope of the risks, and then propose the safeguards needed to mitigate them. Next, if the risks are systemic in nature or concern common platforms, either internal and external, the company's data protection officer must consult with the supervisory authority prior to any processing.

*Fast and broken in the real world.* Facebook, despite having moved away from the aforementioned motto, has continued to be plagued by it. In 2018, it revealed two major breaches: first, that their APIs allowed Cambridge Analytica to illicitly harvest<sup>33</sup> personal data from 87M users, and then their new *View As* feature was exploited<sup>28</sup> to gain control over 50M user accounts. However, this practice of prioritizing speed over security is not limited to one organization. For example, in November 2017, fitness app Strava released an athlete motivation tool called global heatmap that visualized athletic activities of worldwide users. However, within months, these maps were used to identify undisclosed military bases and covert security operations,<sup>27</sup> jeopardizing missions and lives of soldiers.

**Hiding data breaches.** The notion that one is *innocent until proven guilty* predates all computer systems. As a legal principle, it dates back to 6<sup>th</sup> century Roman empire,<sup>3</sup> where it was codified that *prooflies on him who asserts, not on him who denies*. Thus, in the event of a data breach or a privacy violation, organizations typically claim innocence and ignorance, and seek to be absolved of their responsibilities. However, GDPR makes such presumption conditional on the controller proactively

implementing risk-appropriate security measures (that is, accountability), and notifying breaches in a timely fashion (that is, transparency).


ARTICLE 5: PRINCIPLES RELATING TO PROCESSING. “(1) *Personal data shall be processed with ... lawfulness, fairness and transparency; ... purpose limitation; ... data minimization; ... accuracy; ... storage limitation; ... integrity and confidentiality.* (2) *The controller shall be responsible for, and be able to, demonstrate compliance with (1).*”

ARTICLE 33: NOTIFICATION OF A PERSONAL DATA BREACH. “(1) *the controller shall without undue delay and not later than 72 hours after having become aware of it, notify the supervisory authority. ...* (3) *The notification shall at least describe the nature of the personal breach, ... likely consequences, and ... measures taken to mitigate its adverse effects.*”


GDPR’s goal is twofold: first, to reduce the frequency and impact of data breaches, article 5 lays out several ground rules. Controllers are not only expected to adhere to these internally but also be able to demonstrate their compliance externally. Second, to bring transparency in handling data breaches, articles 33 and 34 mandate a 72-hour notification window within which the controller should inform both the supervisory authority and the affected people.

*Data breaches in the real world.* In recent years, responses to personal data breaches have been ad hoc: while a few organizations have been forthcoming, others have chosen to refute,<sup>11</sup> delay,<sup>16</sup> or hide by paying off hackers.<sup>18</sup> However, GDPR’s impact has been swift and clear. Just in the first eight months (May 2018 to Jan 2019), regulators received 41,502 data breach notifications.<sup>12</sup> This number is in stark contrast from the pre-GDPR era, with reports of 945 worldwide data breaches in the first half of 2018.<sup>34</sup>

**Making unexplainable decisions.** Algorithmic decision-making has been successfully applied to several domains: curating media content, managing



**Given the importance of personal data, and the implications of misusing that data, we believe system designers should examine their systems for these anti-patterns, and work toward eliminating them with urgency.**



industrial operations, trading financial instruments, personalizing advertisements, and even combating fake news. Their inherent efficiency and scalability (with no human in the loop) has made them a necessity in modern system design. However, GDPR takes a cautious view of this trend.

ARTICLE 22: AUTOMATED INDIVIDUAL DECISION-MAKING. “(1) *The data subject shall have the right not to be subject to a decision based solely on automated processing ...*”

ARTICLE 15: RIGHT OF ACCESS BY THE DATA SUBJECT. “(1) *The data subject shall have the right to obtain from the controller ... meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing.*”

On one hand, privacy researchers from Oxford postulate<sup>14</sup> that these two regulations, together with recital 71, establish a “right to explanation” and thus, human interpretability should be a design consideration for machine learning and artificial intelligence systems. However, another group at Oxford argues<sup>37</sup> that GDPR falls short of mandating this right by requiring users to demonstrate significant consequences, to seek explanation only after a decision has been made, and to have to opt out explicitly.

*Decision-making in the real world.* The debate over interpretability in automated decision-making has just begun. Starting 2016, the machine learning and artificial intelligence communities began exploring this rigorously: *The Workshop on Explainable AI at IJCAI*, and the *Workshop on Human Interpretability in Machine Learning at ICML* being two such efforts. In January 2019, privacy advocacy group NoYB has filed<sup>23</sup> complaints against eight streaming services including Amazon, Apple Music, Netflix, SoundCloud, Spotify, YouTube, Flimmit, and DAZN for violating article 15 requirements in their recommendation systems.

**Concluding Remarks**

Achieving compliance with GDPR, while mandatory for companies working with




personal data of Europeans, is not trivial. In this article, we examine how the design, architecture, and operation of modern cloud-scale systems conflict with GDPR. Specifically, we illustrate this tussle via six GDPR anti-patterns, that use patterns of system design and operation, which are effective in their own context but violate the rights and regulations of GDPR. Given the importance of personal data, and the implications of misusing that data, we believe system designers should examine their systems for these anti-patterns, and work toward eliminating them with urgency.

**Open issues.** While our list of GDPR anti-patterns is a useful beginning point, it is not exhaustive. Neither have we proposed a methodology for identifying a large number of such anti-patterns, nor do we prescribe any mechanisms toward eliminating them. The six anti-patterns highlighted here exist due to technical and economic reasons that may not entirely be in the control of individual companies. Thus, solving such deep-rooted issues would likely result in significant performance overheads, slower product rollouts, and re-organization of data markets. The equilibrium points of these tussles are not yet clear.

**Future directions.** While there have been a number of recent works analyzing GDPR from privacy and legal perspectives,<sup>5,19,15,35,36,38</sup> the systems community is just beginning to get involved. GDPR compliance brings several interesting challenges to system design. Prominently, addressing compliance at the level of individual infrastructure components (such as, compute, storage, and networking) versus achieving end-to-end compliance of individual regulations (such as, implementing right-of-access in a music streaming service) will result in different trade-offs. The former approach makes the effort more contained and thus, suits the cloud model better. Examples of this direction include GDPR compliant Redis,<sup>30</sup> Compliance by construction,<sup>29</sup> and Data protection database.<sup>20</sup> The latter approach provides opportunities for cross-layer optimizations (for example, avoiding access control in multiple layers). Google search's implementation<sup>2</sup> of Right to be forgotten is in this direction.

Another challenge arises from GDPR being vague in its technical specifications (possibly to allow for technological advancements). Thus, questions like *how soon after a delete request should that data be actually deleted* could be answered in several compliant ways. The idea that compliance could be a spectrum, instead of a well-defined point gives rise to interesting system trade-offs as well as the need for benchmarks that quantify a given system's compliance behavior.

While GDPR is the first comprehensive privacy legislation in the world, several governments are actively drafting and rolling out their own privacy regulations. For instance, California's Consumer Privacy Act (CCPA)<sup>4</sup> went into effect on Jan 1, 2020. We hope that this paper helps all the stakeholders in avoiding the pitfalls in designing and operating GDPR-compliant personal-data processing systems. 

**References**

1. Alexander, F. Data is the new bacon. *IBM Business analytics blog*, 2016; <https://www.ibm.com/blogs/business-analytics/datais-the-new-bacon/>.
2. Bertram, T. et al. trait, A., Thomas, K., and Verney, A. Five years of the Right to Be Forgotten. *ACM CCS*, 2019.
3. Buckland, W. and Stein, P. *A Textbook of Roman Law: From Augustus to Justinian*. Cambridge University Press, 2007.
4. California Consumer Privacy Act. *California Civil Code, Section 1798.100* (Jun 28, 2018).
5. Casey, B., Farhangi, A., and Vogl, R. Rethinking explainable machines: The GDPR's right to explanation debate and the rise of algorithmic audits in enterprise. *Berkeley Technology Law J.* 34 (2019), 143.
6. CNIL. The CNIL's restricted committee imposes a financial penalty of 50 million euros against Google LLC, 2019; <https://www.cnil.fr/en/cnils-restricted-committee-imposes-financial-penalty-50-million-euros-against-google-llc>.
7. Conger, K. How to download your data with all the fancy new GDPR tools. *Gizmodo*, 2018; <https://gizmodo.com/how-todownload-your-data-with-all-the-fancy-new-gdpr-t-1826334079>.
8. Data Deletion on Google Cloud Platform, 2018; <https://cloud.google.com/security/deletion/>.
9. Davies, J. GDPR mayhem: Programmatic ad buying plummets in Europe. *Digiday*, 2018; <https://digiday.com/media/gdpr-mayhem-programmatic-ad-buying-plummets-europe/>.
10. Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. We value your privacy ... Now take some cookies: Measuring the GDPR's impact on Web privacy. *NDSS*, 2019.
11. Doshi, V. 2018. A security breach in India has left a billion people at risk of identity theft. *Washington Post*, 2018; <https://wapo.st/3389hOn>.
12. European Data Protection Board. EDPB: First Year GDPR—taking stock. EDPB News; [https://edpb.europa.eu/news/news/2019/1-year-gdpr-taking-stock\\_en](https://edpb.europa.eu/news/news/2019/1-year-gdpr-taking-stock_en).
13. Forni, A., and Meulen, R. Organizations are unprepared for the 2018 European Data Protection Regulation. *Gartner*, 2017.
14. Goodman, B. and Flaxman, S. European Union regulations on algorithmic decision-making and a right to explanation. *AAAI AI Magazine* 38, 3 (2017).
15. Greengard, S. Weighing the impact of GDPR. *Commun. ACM* 61, 11 (2018), 16–18.
16. Grothaus, M. Panera Bread leaked millions of customers' data. In *Fast Company*, 2018; <https://bit.ly/3cG5jQk>.

17. Haselton, T. Credit reporting firm Equifax says data breach could potentially affect 143 million US consumers. *CNBC*, 2017.
18. Isaac, M., Benner, K., and Frenkel, S. Uber hid 2016 breach, paying hackers to delete stolen data. *New York Times*, 2017; <https://www.nytimes.com/2017/11/21/technology/uber-hack.html>.
19. Koenig, A. Patterns and antipatterns. *J. Object-Oriented Programming* 8, 1 (1995), 46–48.
20. Kraska, T., Stonebraker, M., Brodie, M., Servan-Schreiber, S. and Weitznar, D. DATUMDB: A data protection database proposal. In *Proceedings of Poly'19 co-located at VLDB*.
21. Larson, S. Every single Yahoo! account was hacked—3 billion in all. *CNN Business*, 2017.
22. Lomas, N. Even the IAB warned adtech risks EU privacy rules. *TechCrunch*, 2019; <https://techcrunch.com/2019/02/21/even-theiab-warned-adtech-risks-eu-privacy-rules/>.
23. Lomas, N. Privacy campaigner Schrems slaps Amazon, Apple, Netflix, others with GDPR data access complaints. *TechCrunch*, 2019.
24. Lunden, I. UK's ICO fines British Airways a record £183M over GDPR breach that leaked data from 500000 users. *TechCrunch*, 2019.
25. O'Flaherty, K. Marriott faces £123 million fine for 2018 mega breach. *Forbes*, 2019.
26. OJEU. General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46. *Official Journal of the European Union* 59 (2016), 1–88.
27. Quarles, J. An update on the global heatmap, 2018; <https://blog.strava.com/press/a-letter-to-the-strava-community/>.
28. Rosen, G. Security Update, 2018; <https://newsroom.fb.com/news/2018/09/security-update/>.
29. Schwarzkopf, M., Kohler, E., Kaashoek, F., and Morris, R. GDPR compliance by construction. In *Proceedings of Poly'19 co-located at VLDB*.
30. Shah, A., Banakar, V., Shastri, S., Wasserman, M., and Chidambaram, V. Analyzing the impact of GDPR on storage systems. *USENIX HotStorage*, 2019.
31. Shastri, S., Banakar, V., Wasserman, M., Kumar, A., and Chidambaram, V. Understanding and benchmarking the impact of GDPR on database systems. In *Proceedings of the VLDB Endowment* 13, 7 (2020).
32. Shastri, S., Wasserman, M., and Chidambaram, V. The seven sins of personal-data processing systems under GDPR. *USENIX HotCloud*, 2019.
33. Solon, O. Facebook says Cambridge Analytica may have gained 37M more users' data. In *The Guardian*, 2018; <https://bit.ly/2S9xVYF>.
34. Targett, E. 6 Months, 945 Data Breaches, 4.5 Billion Records. *Computer Business Review*, 2018; <https://www.cbronline.com/news/globaldata-breaches-2018>.
35. Tesfay, W., Hofmann, P., Nakamura, T., Kiyomoto, S., and Serna, J. I read but don't agree: Privacy policy benchmarking using machine learning and the EU GDPR. In *Companion Proceedings of the Web Conference*, 2018, 163–166.
36. Utz, C., Degeling, M., Fahl, S., Schaub, F., and Holz, T. (Un) informed consent: Studying GDPR consent notices in the field. *ACM CCS*, 2019.
37. Wachter, S., Mittelstadt, B., and Floridi, L. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *Intern. Data Privacy Law* 7, 2 (2017), 76–99.
38. Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law & Technology* 31 (2017), 841.

**Supreeth Shastri** is an assistant professor of computer science at the University of Iowa, IA, USA.

**Melissa Wasserman** is the Charles Tilford McCormick Professor of Law at the University of Texas at Austin, TX, USA.

**Vijay Chidambaram** is an assistant professor of computer science at the University of Texas at Austin, TX, USA.

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

DOI:10.1145/3382037

## An approach to reproducibility problems related to porting software across machines and compilers.

BY DONG H. AHN, ALLISON H. BAKER, MICHAEL BENTLEY, IAN BRIGGS, GANESH GOPALAKRISHNAN, DORIT M. HAMMERLING, IGNACIO LAGUNA, GREGORY L. LEE, DANIEL J. MILROY, AND MARIANA VERTENSTEIN

# Keeping Science on Keel When Software Moves

HIGH PERFORMANCE COMPUTING (HPC) is central to solving large problems in science and engineering through the deployment of massive amounts of computational power. The development of important pieces of HPC software spans years or even decades, involving dozens of computer and domain scientists. During this period, the core functionality of the software is made more efficient, new features are added, and the software is ported across multiple platforms. Porting of software in general involves the change of compilers, optimization levels, arithmetic libraries, and many other aspects that determine

the machine instructions that actually get executed. Unfortunately, such changes do affect the computed results to a significant (and often worrisome) extent. In a majority of cases, there are not easily definable a priori answers one can check against. A programmer ends up comparing the new answer against a trusted baseline previously established or checks for indirect confirmations such as whether physical properties such as energy are conserved. However, such non-systematic efforts might miss underlying issues, and the code may keep misbehaving until these are fixed.

In this article, we present real-world evidence to show that ignoring numerical result changes can lead to misleading scientific conclusions. We present techniques and tools that can help computational scientists understand and analyze compiler effects on their scientific code. These techniques are applicable across a wide range of examples to narrow down the root-causes to single files, functions within files, and even computational expressions that affect specific variables. The developer may then rewrite the code selectively and/or suppress the application of certain optimizations to regain more predictable behavior.

Going forward, the frequency of required ports of computational software will increase, given that performance gains can no longer be obtained by merely scaling up the clock frequency, as used to be possible in prior decades. Performance gains are now hinged on the use of multicore CPUs, GPUs and other accelerators, and above all, advanced compilation methods. While reproducibility

### » key insights

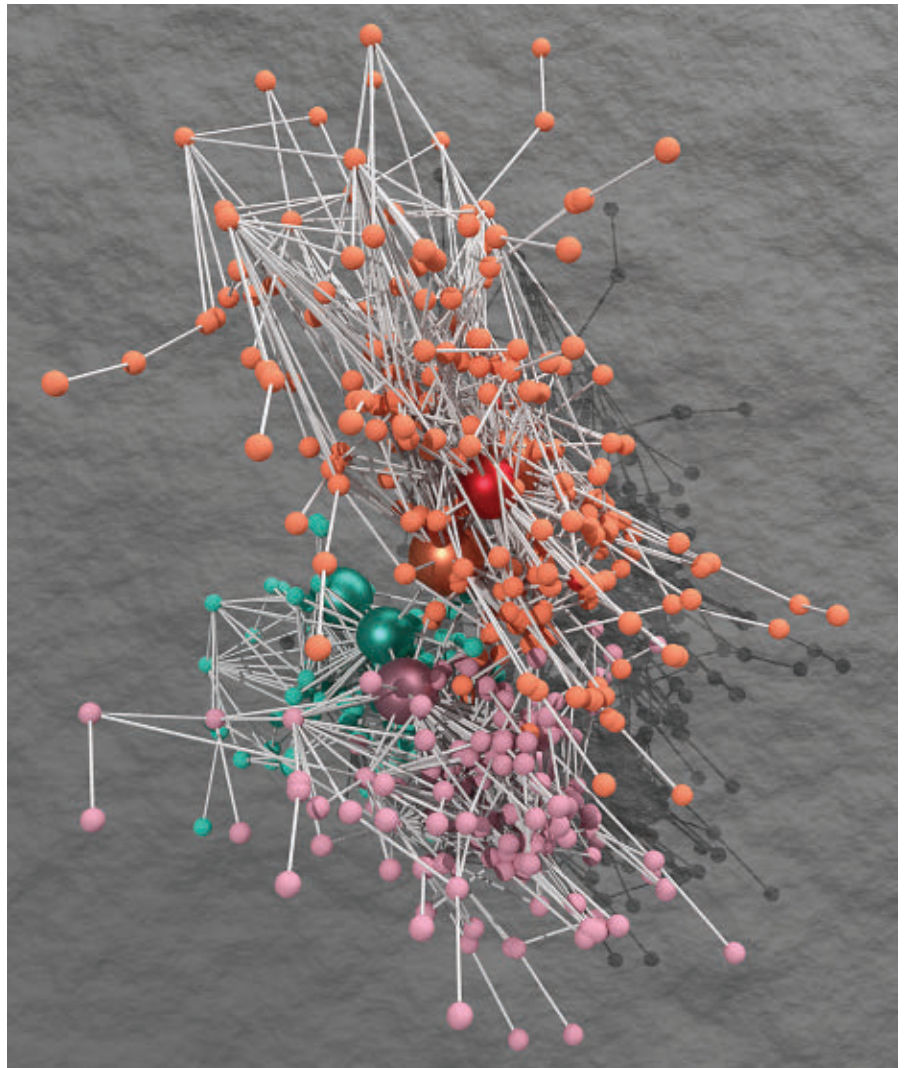
- Even seemingly small changes to scientific software or its build and runtime environment can create large, unexpected changes in floating-point results.
- Bisection search helps locate code sites that sow variability, often with firm guarantees.
- Combining statistical consistency testing with graph-based code analysis provides important insight into sources of floating-point variability in the Community Earth System Model (CESM™).

across compilers and platforms in this sense is a problem that hasn't grabbed headlines in discussions centered around reproducibility, the problem is real (see sidebar "Is There a Reproducibility Problem?") and threatens to significantly affect the trustworthiness of critical pieces of software.

It may seem that all the difficulties described thus far can be solved by ensuring that compilers adhere to widely accepted rigorous standards of behavior spanning machines and optimization levels. Unfortunately, this goal is extremely difficult to realize in principle as well as in practice. Modern compilers must exploit not only advanced levels of vectorization but also the characteristics of heterogeneous computing platforms. Their optimizations in this complex space are triggered differently—even for the same compiler flags—based on the compiler's projection of the benefits of heeding the flags. This behavior is very difficult to characterize for all cases. While vendor compilers are often preferred for their superior performance—especially with respect to vectorization—they also present a challenge in terms of intervention in case issues arise.

In this article, we describe the extent of this challenge, and what is actionable in terms of equipping developers with practical tools (FLiT, CESM-ECT, and CESM-RUANDA). Some of these tools are already usable today for important codes such as hydrodynamics simulation codes and finite element libraries. We then take up the more challenging problem of climate simulation codes where much more work is needed before an adequate amount of tooling support will be developed. We describe the progress already made in this area by describing our solutions that address Earth system models (ESMs) that are central to climate simulation.

**"Climate-changing" compiler optimizations.** Earth system models (ESMs) simulate many physical, chemical, and biological processes and typically feature a complex infrastructure that couples separate modular representations of Earth system components (for example,



**A three-dimensional, undirected representation of the example from Figure 6. Nodes are colored by community membership and sized based on a threshold centrality value. The red nodes represent model variables sensitive to specific CPU instructions. All nodes with eigenvector centrality  $\leq 0.4$  have a constant size, and those above the threshold are scaled and highlighted by increased reflectance. Credit: Liam Krauss of LLNL.**

atmosphere, ocean, land, river, ice, and land ice). ESMs are characterized by exceedingly large code bases that have resulted from decades of development, often containing a mix of both legacy code and more modern code units. Further, most ESMs are in a state of near constant development as advancing scientific discovery requires the continual addition of new features or processes, while rapidly evolving HPC technology requires new optimizations of the code base. Needless to say, software engineering for ESMs is challenging,

and quality assurance is particularly critical for maintaining model credibility given that output may have policy and societal impact as future climate scenarios are considered.<sup>7,10,23</sup>

The popular Community Earth System Model (CESM<sup>TM</sup>)<sup>13</sup> is a fully coupled community global climate model that enjoys widespread use across a range of computational platforms, including cutting-edge HPC architectures. With a code base of nearly two million lines across approximately 13,000 subroutines and 3,000 functions, it is



## Is There a Reproducibility Problem?

- ▶ In the Community Earth System Model (CESM™) software, the compiler introduced fused multiply add (FMA) instructions that resulted in “climate changing” differences from the baseline simulations.<sup>3</sup>
- ▶ Compiling Laghos (<https://github.com/CEED/Laghos>), a hydrodynamics simulation, under the IBM compiler `xlc` with optimization level `-O3`, there were negative densities created and energy was not conserved after just one iteration.<sup>5</sup>
- ▶ FLIT-based testing of the MFEM finite element library revealed that even reasonable compiler optimization levels can change the result by as much as 190%.<sup>5</sup>

## Statistical Ensemble Consistency Testing

When a climate simulation code is ported to a new platform, the output on the new platform will not be bit-identical to the original. This difference makes answering the question of consistency non-trivial. Instead, we ask a more tangible question: *Is the new output statistically distinguishable from the original?*

The CESM Ensemble Consistency Test (CESM-ECT) was developed to answer this new question. Ensemble methods are common in climate studies, as a collection of simulations are needed to describe the internal variability in the climate model system. (Climate models are inherently chaotic, meaning that even tiny perturbations or differences can cause large effects.) CESM-ECT generates a large “baseline” ensemble on a trusted machine and software stack and utilizes a testing framework based on the popular technique of Principal Component Analysis (PCA) to determine whether a set of new simulations (for example, from a new machine, compiler upgrade, optimization, and so on) is statistically distinguishable from the baseline ensemble. This ensemble-based approach to verification serves as a powerful classification tool when bit-identical requirements are too restrictive.

critical to ensure that changes made during the CESM development life cycle do not adversely affect the model results. A CESM simulation output is only bit-reproducible when the exact same code is run using the *same* CESM version and parameter settings, initial conditions and forcing data, machine, compiler (and flags), MPI library, and processor counts, among others. Unfortunately, control over these quantities to this degree is virtually impossible to attain in practice, and further, because the climate system is nonlinear and chaotic, even a double-precision roundoff-level change will propagate rapidly and result in output that is no longer bit-identical to the original.<sup>21,a</sup> As an example, a port of CESM to a new architecture is a common occurrence that perturbs the model’s calculations (all of which are carried out in

double-precision) and requires an evaluation for quality assurance. While the output on a new machine will not be bit-identical, one would reasonably expect there to be some degree of consistency across platforms, as the act of porting should not be “climate-changing.” We would expect the same scientific conclusions to be reached when analyzing output from model runs that were identical in all but compute platform.

In the past, such CESM consistency checks were costly undertakings that required climate science expertise and multi-century simulations, as there is not a simple metric for what defines climate changing. However, statistical testing techniques have recently been developed that define consistency in terms of statistical distinguishability, leading to the creation of the CESM Ensemble Consistency Test (ECT)<sup>1,2,21</sup> suite of tools (see the sidebar “Statistical Ensemble Consistency Testing”). The simple and efficient CESM-ECT tools are regularly used by CESM software engineers for evaluating ports to new ma-

chines, software upgrades, and modifications that should not affect the climate. In practice, CESM-ECT has proven effective in exposing issues in the CESM hardware and software stacks, including large discrepancies caused by fused multiply-add (FMA) optimizations, an error in a compiler upgrade, a random number generator bug specific to big-endian machines, and an incorrect input parameter in a sea ice model release. In addition, by relaxing restrictive bit-identical requirements, CESM-ECT has allowed greater freedom to take advantage of optimizations that violate bit reproducibility but result in statistically indistinguishable output. Note that optimizing performance for climate models has long been of interest due to their computational expense. For example, a fully coupled “high-resolution” CESM simulation (that is, atmosphere/land at 0.25° grid spacing and ocean at 0.1°) can easily cost on the order of 250,000 core hours per simulated year.<sup>28</sup> While lower resolution simulations consume fewer core hours per simulated year (a 1.0° grid costs  $\approx 3,500$  core hours), these simulations are often run for a large number of years. For example, CESM’s contribution to the current Coupled Model Comparison Project (Phase 6)<sup>11</sup> (used by the Intergovernmental Panel on Climate Change<sup>15</sup> for their assessment reports) is expected to consume nearly 125 million core hours.

### Flitting Behaviors

Compiler optimizations do have the capability to change the result of floating-point computations. However, it is possible, even likely, that these optimizations can generate an answer closer to the scientist’s underlying model. Unfortunately, in general, it is hard to know which of two answers is better. Therefore, the best we can do is to try to reproduce a trusted implementation on trusted hardware. Thus, we focus on reproducibility and consistency of the program’s output compared to the baseline generated from the trusted configuration.

It is clear that manual testing to locate the absence of reproducibility does not scale: any subset of the software submodules could be responsible for the observed result change. Projects that maintain rigorous unit testing may

a Bitwise reproducibility is a coveted goal in general (not just for CESM), as it greatly facilitates regression testing.

already be able to utilize them to locate some problems, however many large projects have insufficient unit testing. Furthermore, floating-point rounding is non-compositional: decreased error in one component can sometimes increase the overall roundoff error.<sup>18,29</sup> It violates some of the basic algebraic laws such as associativity (See the sidebar “Floating-point Arithmetic and IEEE”).

Sources of floating-point behavioral changes are also too numerous. Sometimes hardware implementations have fewer capabilities, such as not supporting subnormal numbers in their floating-point arithmetic.<sup>14</sup> Some strange behaviors can be observed when subnormal numbers are abruptly converted to zero. Other times, there are additional hardware capabilities the compiler may utilize, such as replacing a multiply and an add with a single FMA instruction. While FMA can reduce floating-point rounding error locally (because there is only one rounding step instead of two), care must still be taken. A lower local error does not necessarily equate to lower global error, particularly for a code that is sensitive to roundoff.

Under heavy optimizations, compilers can change the associativity of arithmetic operations such as reductions (especially when code is vectorized). For example, an arithmetic reduction loop whose trip-count is not an integral multiple of the vector lane width must involve an extra iteration, handling the remaining elements. The manner in which this iteration is incorporated can change overall associativity. Given the increasing use of GPUs and other accelerators, one must take into account how they deviate from IEEE floating-point standard in an increasing number of ways. The use of mixed-precision arithmetic where later iterations change precision<sup>6,20,27</sup> can exacerbate all these behaviors.

When a simulation code is affected by any one of these reasons and the computational results are deemed unacceptable, how does a developer proceed? The first step would typically be to find the source(s) of floating-point divergence and try to narrow down the root-causes based on one’s best guess or experience. Next, it seems logical to identify the sites and involved variables that play a part in the numerical inconsistency. Once inconsistent configurations and the associated code sites

are identified, there may be many approaches that can be used to mitigate the inconsistency. For example, one could employ numerical analysis techniques to improve the stability of the underlying algorithm; compile the affected units with fewer optimizations; or, rewrite the units to behave similarly under the two different configurations.

Here, we present a collection of techniques that can be used on realistic HPC codes to investigate significant differences in calculated results.

### FLiT: Tool for Locating Sources of Variability

FLiT is a tool and testing framework meant to analyze the effect of compilers and optimizations on user code. It allows users to compare the results between different compilers and optimizations, and even locate the code sites to the function level where compilation differences cause results to differ.

*Logarithmic search.* Suppose the code is contained in a collection of  $N$  files and a new compilation produces

inconsistent results. We cannot know there is only one variability site or that errors are not canceled out in strange ways. To make any progress, we make the assumption that floating-point differences are unique (for example, no two variability sites exactly cancel out each other). Without this assumption, to be sure we found all variability sites, it would require an exponential search. With this assumption, we can utilize Delta Debugging<sup>30</sup> with complexity  $O(N \log N)$ . However, in practice, we have found most variability sites to act alone, meaning they contribute variability by themselves and not in concert with other components. We then make a further assumption that each site acts alone in contributing variability (call this the singleton assumption). This assumption allows for an efficient logarithmic search as illustrated in Figure 1 with complexity  $O(k \log N)$  where  $k$  is the number of variability sites. Speedometers are also displayed in Figure 1 to represent performance of our partially optimized executable, demon-

## Floating-Point Arithmetic and IEEE

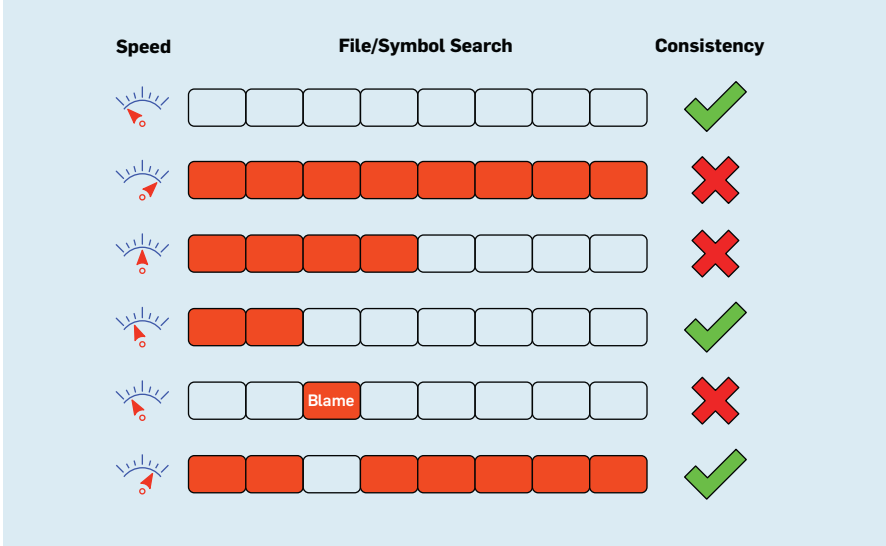
Under most circumstances FMA yields more accurate results than a multiplication and addition computed separately. However, this need not be the case. For example, given the expression  $a * b + ac$ , one expects the expression to evaluate to 0 when  $b = -c$ . However, with FMA, the calculation carried out might be  $a * c + ac$  where  $ac$  represents the result of  $a * c$  with rounding. Kahan<sup>17</sup> presents another example where the multiplication of a complex number by its complex conjugate using FMA might not produce a real number.

## Climate Models Are Important

Given recent warming trends and increases in extreme events, understanding present, past, and future climate scenarios is increasingly a global priority. Models such as CESM that perform state-of-the-science climate simulations are particularly vital for addressing otherwise intractable “what if?” climate questions (for example, “What if all of the ice in the Arctic melts?” or “What if ocean temperatures rise by  $N$  degrees?”), enabling better societal preparation for the future.

CESM 2.0 was released in the summer of 2018, and its popularity is a result of collaborations over several decades between scientists at the National Center for Atmospheric Research (NCAR) and various universities and research institutions. CESM is a true community model that is accompanied by a robust, extensible and portable workflow and code base that provides users a standard way to readily create model experiments and customize the experimental setup. The infrastructure allows users to easily explore and evaluate proposed science changes by creating simplified model configurations (for example, via lower resolutions or disabled feedback). Climate models are well known for pushing the limit for what is computationally feasible, and CESM’s infrastructure permits the extensive testing of the model, thereby ensuring its reliability and efficiency on a broad spectrum of modern computational platforms. Establishing the trustworthiness of a code like CESM is paramount given its critical role in exploring important climate questions and defining consistency separately from bit-reproducibility is a practical necessity.

**Figure 1. Example of the Bisect logarithmic search where shaded blocks represent optimized files or symbols. Unshaded blocks are from the trusted baseline compilation.**



strating that the more files are optimized, the more performance it yields.

The logarithmic search in Figure 1 proceeds as follows. With all of code optimized (all the rectangles shaded), the computation runs quite fast (the speedometer is at its highest), but the results are inconsistent. Even with the left half optimized, the result is still inconsistent. Logarithmic search subdivides the left half, keeping the first two files of the left half optimized, which results in consistency. We then divide the remaining two files from the left half to test file 3 by itself. This file optimized by itself causes inconsistency and is therefore given blame. Removing file 3 from the search, we start over. In this case, we see optimizing all except for file 3 obtains consistency, therefore we have found all sites.

We framed this problem in terms of files, but after blaming files, we can perform this search again over symbols in each file (representing individual functions). The algorithm is the same

but the implementation for symbols is a bit more complicated, as outlined in Bentley et al.<sup>5</sup>

*Verifying the singleton assumption.* A check is inserted in the search that probably verifies whether the singleton assumption holds.<sup>5</sup> In fact, as shown in this illustration, it may be possible to judiciously add back some units in an optimized mode (the last row from Figure 1) to finally leave the code highly optimized and producing acceptable answers. It would also be advantageous to obtain an overall speedup profile of one’s simulation code. One such profile can be seen in Figure 3. This was obtained for an example supplied with a widely used finite element library, MFEM. From this profile, one can observe that it is possible to attain a speedup of 9.4% (compared with `gcc -O2`) with exact reproducibility, or a speedup of 39.6% with a small amount of variability.

*FLiT workflow.* The FLiT workflow is shown in Figure 2. A full application or a piece of it may be converted into a

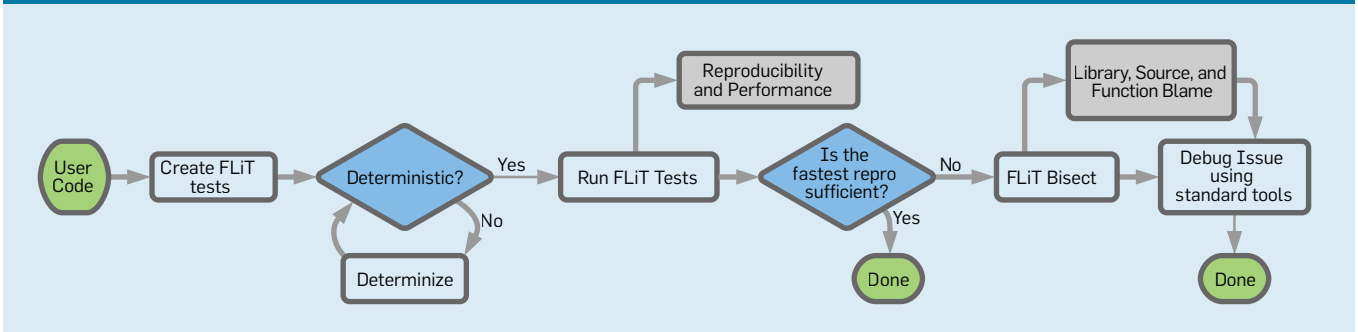
FLiT test. The given FLiT test, sequential or parallel (OpenMP or MPI), must be run-to-run deterministic. One must attempt to make their code as deterministic as possible before using FLiT. For example, random number generators can be seeded and MPI applications can use capture-playback (using tools like ReMPI<sup>24</sup>).

The FLiT test can now be compiled in various ways and run to find the compilations that cause significant differences. If one of the compilations delivers results within tolerance and has acceptable performance, the configuration search can end. For example, in Figure 3, we obtain a 9% speedup with a bitwise equal result on MFEM example 9, and if some variability can be tolerated, then the compilation with 40% speedup can be used. But when significant speedups are accompanied with unacceptable differences, the FLiT Bisect search can be used to locate the sites of variability. The FLiT Bisect search proceeds as previously described.

FLiT is a publicly released tool.<sup>4</sup> It has been applied to production codes within the Lawrence Livermore National Laboratory (LLNL) and has successfully located issues in the MFEM library and the Laghos application, as described earlier. FLiT Bisect first performs *File Bisect*, which proceeds as follows:

1. compile each source file into an object file using the trusted baseline compilation, and another object file using the optimization compilation under test.
2. get the next file combination to try from the logarithmic search.
3. link together the chosen object files from the two compilations to make a single executable (see File Bisect in Figure 4).
4. run this generated executable and

**Figure 2. FLiT workflow.**





compare with the baseline run results.

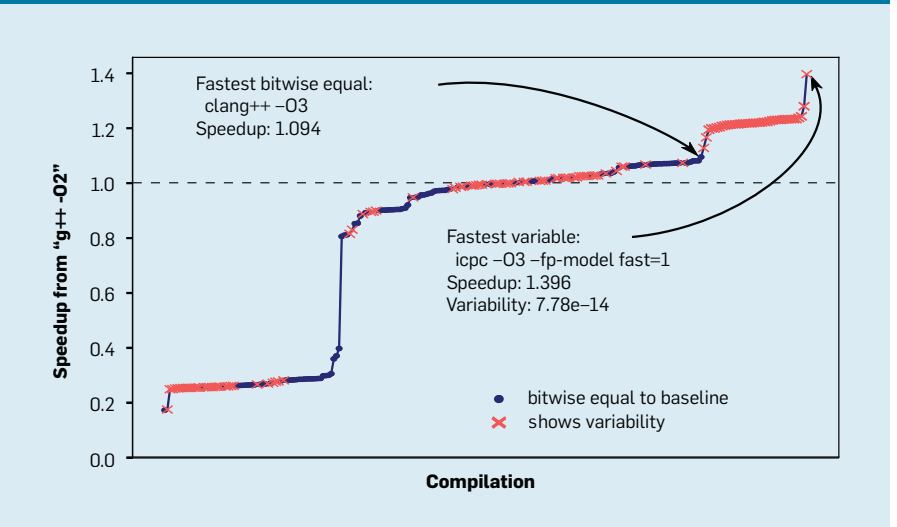
5. return the comparison to the search algorithm and repeat from (2).

The runtime of FLiT Bisect is the time it takes to run the test code times the number of file combinations and symbol combinations to be evaluated. Notice that compilation into object files happens only at the beginning. After that, FLiT simply does a link step and run for each search step. It is worth noting that FLiT Bisect also includes the capability to report how much each site is estimated to contribute to the overall result divergence.

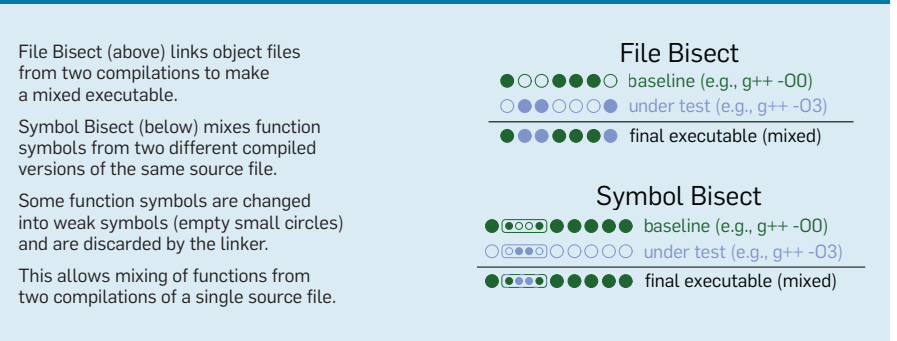
*Function-level Bisect.* While File Bisect is quite useful in narrowing down the reasons for a software’s non-portability, we often have to locate problems at a finer level of resolution—meaning, a single function within a file. FLiT supports this via its *Symbol Bisect* feature. As seen in Figure 4, Symbol Bisect mixes compiled functions from two different compilations of the same source file. This is performed by demoting some symbols to be weak symbols. During link-time, if there is a duplicate symbol but one is weak, then the strong symbol is kept while the weak symbol is discarded. This approach allows FLiT to search over the symbol space *after optimizations have been performed*. However, for this to be effective, the `-fPIC` compilation flag must be used (only on the object file to be mixed) to ensure no inlining between functions that we might want to replace occurs. FLiT checks whether using `-fPIC` interferes with the optimization that causes the result difference.

In practice, this modality of search has helped us successfully attribute root causes down to a small set of functions. For example, in the case of Test-13 within the MFEM library, FLiT-based testing revealed that a compiler optimization level that involved the use of AVX2, FMA, and higher precision intermediate floating-point values produced a result that had a relative difference of 193% from the baseline of `g++ -O2`. The  $L_2$  norm over the mesh went from approximately 5 to 15 after the optimizations. Using Symbol Bisect, the problem was located to be within one simple function that calculates  $M = M + aAA^T$ , with  $a$  being a scalar, and  $M$  and  $A$  being dense square matrices. This case wasn’t known to the developers of MFEM.

**Figure 3. Performance profile of compilations of Example 9 from MFEM. The compilations with the fastest bitwise equal and fastest overall speeds are labeled.**



**Figure 4. File Bisect and Symbol Bisect.**



Conversation with the developers of MFEM is under way to resolve this issue. This finding may indicate numerical instability of the underlying finite element method employed, or with its implementation.

Addressing the identified and located issue is outside of FLiT’s scope. It is then the responsibility of the scientific software developer to solve the issue in order to obtain consistency and numerical stability. A designer may then choose to solve the identified non-portability either by tuning precision, rewriting the computation differently (perhaps employing more numerically stable approximations), or avoiding the problematic optimization for the whole application or the affected files.

**CESM**

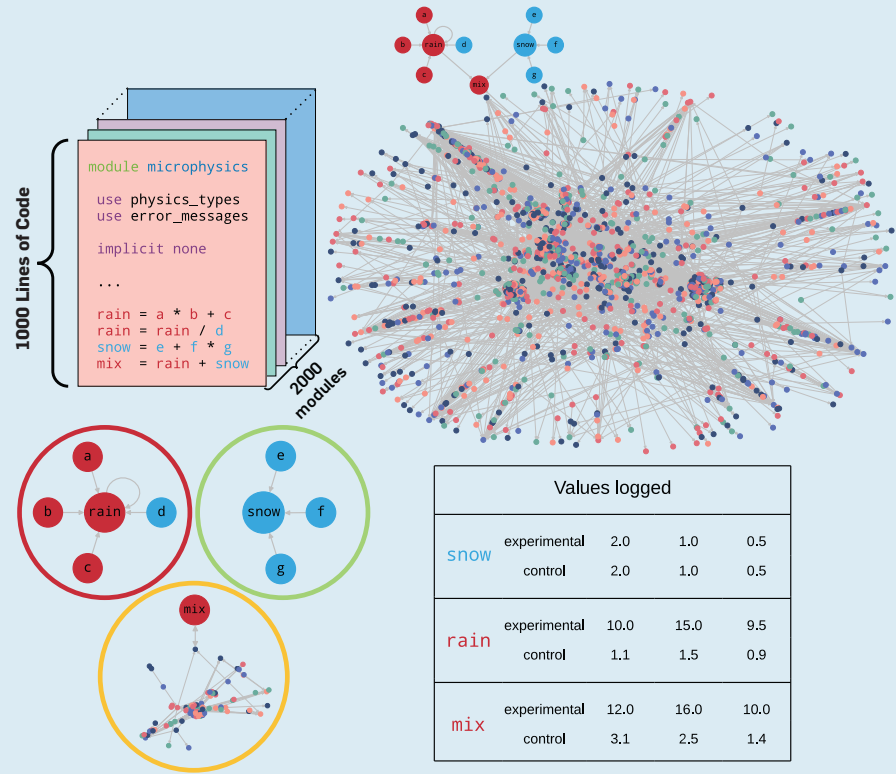
FLiT’s tolerance-based approach to consistency will work for many code bases, but for applications that model complex and chaotic systems, a more nuanced method may be needed. For

CESM, statistical consistency between a baseline ensemble and a set of new runs is determined by the CESM-ECT quality assurance framework. Extending the CESM-ECT to help understand why new runs are inconsistent is crucial for comprehensive quality assurance for CESM. Retaining in mind our long-term goal of impacting other large, critical applications, we now describe our recent efforts to tackle the challenge of root cause analysis of inconsistency in CESM.

The CESM-ECT has proven to be useful in terms of detecting inconsistencies that were either introduced during the process of porting the CESM software or by a new machine platform itself, both of which are not uncommon. Such sources of inconsistency can be true errors (for example, resulting from a compiler bug) or new machine instructions. However, while CESM-ECT issues a “fail” when a statistical discrepancy is identified in the new output, little useful information is provided about the possible cause.

**Figure 5. The CESM root cause identification workflow.**

In this example, the computation of the fictitious CESM output variable rain causes CESM-ECT failure due to an operation in the notional "microphysics" module. To find the cause, CESM-RUANDA first converts the CESM source code (top left) to a directed graph (top right). The subgraph responsible for computing rain is partitioned into communities (bottom left). Then nodes within the communities are selected for runtime sampling by their centrality. The table (bottom right) illustrates how runtime value comparison between an experimental and control case at three logged execution points (columns) can reveal the cause of the CESM-ECT failure.



This lack of fine-grained information can be quite frustrating for the user, who would like to know why the new run failed so that the problem can be addressed. And while debugging a large and complex code like CESM is challenging in general, some hope generally exists when the code crashes or stalls or the numerics blow up. In these situations, we often have enough information (from a large-scale debugging tool or software output) to roughly determine the source of the error. However, when trying to determine the cause of a statistical discrepancy in CESM output, it may be far from clear where (or even how) to start looking for the root cause.

*Automating root cause analysis for CESM.* The need for an automated tool that enables developers to trace a problem detected in CESM output to its source was felt acutely shortly after CESM-ECT was first put into use for verifying ports to other platforms (against simulations on the NCAR supercomputer). Only one of many CESM-supported platforms failed the CESM-ECT and determining the cause of the failure took several frustrating months of effort

from a number of scientists and engineers to identify FMA instructions as giving rise to inconsistency (for example, see Baker et al.<sup>3</sup>). Ideally, a companion tool to CESM-ECT would identify which *lines of code* or *CPU instructions* were responsible for the failure. While tools do exist to find differences at this level, we were not aware of any that we could directly apply to a code the size and complexity of CESM. Approaches based on SAT or Satisfiability Modulo theories are precise, but often cannot handle large code bases.<sup>25</sup> Debugging and profiling toolkits are capable of detecting divergent values in individual variables, but the sampling process can be expensive as well. Furthermore, identifying which variables to sample is a formidable challenge. Therefore, we adopted the strategy of reducing the search space for the root cause(s) to a tractable quantity of code that would facilitate the use of tools like FLiT or KGEN<sup>19</sup> or runtime sampling.

We have successfully progressed toward our goal via a series of developed techniques that we collectively refer to as the CESM Root caUse Analysis of Numerical Discrepancy (CESM-RUAN-

DA).<sup>22</sup> This toolkit parses the CESM source code and creates a directed graph of internal CESM variables that represents variable assignment paths and their properties. Based on its determination of which CESM output variables are most affected (using information from CESM-ECT), it then extracts a subgraph responsible for calculating the output variables via a form of hybrid program slicing. Next, the subgraph is partitioned into communities to facilitate analysis, and nodes are ranked by information flow within the communities using centrality. The centrality-based ranking enables either runtime sampling of critical nodes or the identification of critical modules that can be individually extracted from CESM and run as an independent kernel (for example, via KGEN). See Figure 5 for a visual depiction of CESM-RUAN-DA. Translating the CESM source code into a directed graph representation enables fast, hybrid analysis of information flow making it easier for other existing tools or techniques to locate problematic lines of CESM code.

As an example, CESM-RUAN-DA can identify internal CESM variables whose

values change markedly when computed with FMA. CESM built by the Intel 17 compiler with FMA enabled generates output on the NCAR supercomputer that is flagged as a failure by CESM-ECT. After pinpointing the output variables most affected by enabling FMA instructions, CESM-RUANDA narrows the root cause search space to a subgraph community corresponding to the model atmosphere microphysics package. Examining the top nodes ranked by centrality yields several of the internal variables that take very different values with FMA enabled (Figure 6), allowing us to reach the same conclusion as the manual investigation into the failing CESM port in a fraction of the time (less than an hour on a single CPU socket). The automated identification of the root causes of discrepancies detected in CESM output provided by CESM-RUANDA will tremendously benefit the CESM community and developers.

It is important to highlight that while a CESM-ECT “fail” has a negative connotation, it is simply an indicator of statistically differentiable output. While the negative connotation is warranted for bugs, it masks a subtlety in the case of FMA. In keeping with the sidebar on floating-point arithmetic, we note that CESM-ECT does not indicate which output (with

FMA or without) is more “correct” (in terms of representing the climate state). While domain experts might be able to make such a determination, the model should ideally return consistent results regardless of whether FMA machine instructions are executed. In this case, our tools seem to indicate an instability or sensitivity in portions of the code that ideally could be corrected with a redesign.

### Concluding Remarks

Computational reproducibility has received a great deal of (well-deserved) attention, with publications emphasizing the reproducibility of experimental methods in systems<sup>8</sup> through summaries of workshops covering scientific and pragmatic aspects of reproducibility.<sup>16</sup> While the problems due to non-reproducibility are amply clear, there is a dearth of tools that help solve day-to-day software engineering issues that impact software developers as well as users.

In this context, our specific contribution in this paper has been a two-pronged approach that allows domain scientists to act on reproducibility problems related to porting software across machines and compilers. Our first specific contribution is FLiT—a tool that can be applied to real-world libraries and applications when they exhibit non-

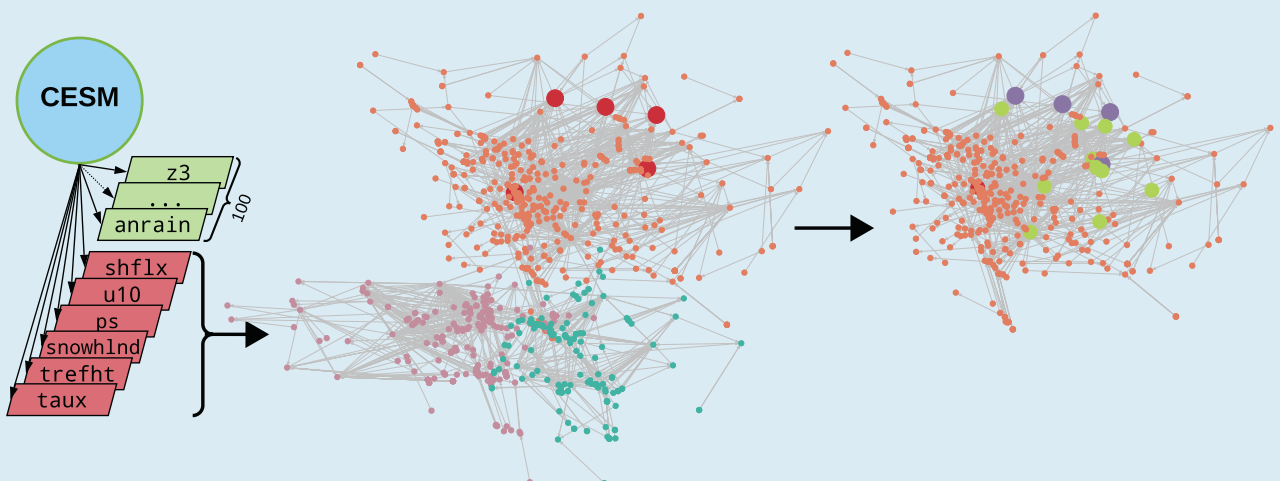
reproducible behavior upon changing compilers or optimization flags. Our second contribution are the CESM-ECT and CESM-RUANDA tools that have approached the problem on a very large scale and in the context of climate simulation software.

While much more work remains to be done on both tools, the anticipated usage model is to first use CESM-ECT to determine that a discrepancy exists, then employ CESM-RUANDA to narrow down the scope of the problem (in codebases exceeding several million lines of code) to specific variables whose values differ significantly, and finally attribute the root cause to individual files or functions via tools such as FLiT. The efficacy of FLiT was demonstrated on the MFEM code which occupies over 100K lines of code and consists of 2,998 functions spread over 97 source files. With such non-trivial code sizes already handled via FLiT, we believe that a combination of these tools will quite naturally lead to an overall superior diagnostic process.

*Building a community is essential.* To help increase the list of tools and approaches in this area, we are eager to engage in collaborations in two primary directions. First, the FLiT tool is available publicly at <https://github.com/PRUNERS>. We are open to developing

**Figure 6. A schematic representation of CESM-RUANDA applied to the problem of finding variables most affected by FMA instructions.<sup>22</sup>**

Of the more than 100 atmosphere output variables used in the CESM-ECT, six are related to the failure (left). The CESM subgraph that computes these six variables is represented by the center plot, where node color designates community membership. Note that we render a smaller subgraph than that produced in Milroy et al.<sup>22</sup> for illustrative purposes. The large red nodes in the center plot represent five variables most affected by FMA instructions. In the rightmost plot the community containing these five variables is isolated and nodes are selected for runtime sampling by their centrality. Large green nodes are those chosen for sampling and purple nodes are variables sensitive to FMA which are also selected for sampling. All but one red node from the center plot would be identified by CESM-RUANDA.





FLIT with external input, collaborations, and feature requests. Second, ideas centered around the CESM-RU-ANDA are ripe for re-implementation, and at NCAR, we are open to supplying computational kernels from the publicly available CESM code to the community. The ideas as well as results behind FLIT and CESM-RUANDA are described in greater detail in Bentley et al.<sup>5</sup> and Milroy et al.,<sup>22</sup> respectively. To further help with community building, we have recently contributed a collection of open-source tools as well as conference tutorials that help pursue many of the issues surrounding floating-point precision analysis, tuning, and exception handling; these are available for perusal at <http://fpanalysistools.org>.<sup>12</sup>

In summary, the integrity of computational science depends on minimizing semantic gaps between the source level representation of simulation software and its executable versions. Such gaps arise when hardware platforms change, libraries change, and compilers evolve. These changes are necessitated by the need to maintain performance in the present post Dennard scaling era. Furthermore, the pace of these changes is only bound to increase as the designer community is highly engaged in squeezing out the last drop of performance from current generation (as well as upcoming) machines and runtimes. Therefore, the onus of computer science researchers is not only to minimize or avoid these gaps through formally verified compilation methods (for example, CompCert<sup>9</sup>), develop tools that discover and bridge these gaps, and also make fundamental advances that contribute to reproducibility (for example, recent contributions to the IEEE-754 standard in support of reproducible arithmetic operations.<sup>26</sup>).

**Digital content available for inclusion with this article.** Sources and detailed instructions to install and use the FLIT software system on a worked-out example of debugging a scenario within the MFEM finite element library is available from <http://fpanalysistools.org>.

**Acknowledgments.** This work was performed under the auspices of the U.S. Department of Energy by LLNL under contract DE-AC52-07NA27344 (LLNL-CONF-759867), and supported

by NSF CCF 1817073, 1704715. The CESM project is supported primarily by the National Science Foundation (NSF). This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the NSF under Cooperative Agreement No. 1852977. Computing and data storage resources, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. **C**

**References**

1. Baker, A.H. et al. A new ensemble-based consistency test for the community earth system model. *Geoscientific Model Development* 8, 9 (2015), 2829–2840; doi:10.5194/gmd-8-2829.
2. Baker, A.H. et al. Evaluating statistical consistency in the ocean model component of the Community Earth System Model (pyCECT v2.0). *Geoscientific Model Development* 9, 7 (2016), 2391–2406; https://doi.org/10.5194/gmd9-2391-2016.
3. Baker, A.H., Milroy, D.J., Hammerling, D.M., and Xu, H. Quality assurance and error identification for the Community Earth System Model. In *Proceedings of the 1st Intern. Workshop on Software Correctness for HPC Applications*. ACM, New York, NY, USA, 8–13; https://doi.org/10.1145/3145344.3145491.
4. Bentley M. and Briggs, I. FLIT Repository, 2019; https://github.com/PRUNERS/FLIT.git
5. Bentley, M. et al. Multi-level analysis of compiler-induced variability and performance trade-offs. In *Proceedings of the 28th Intern. Symp. High-Performance Parallel and Distributed Computing*. ACM, 2019, 61–72; https://doi.org/10.1145/3307681.3325960
6. Chiang, W-F, Baranowski, M., Briggs, I., Solovoyev, A., Gopalakrishnan, G. and Rakamaric, Z. Rigorous floating-point mixed-precision tuning. In *Proceedings of the 44th ACM SIGPLAN Symp. Principles of Programming Languages*. G. Castagna and A.D. Gordon (Eds.), (Paris, France, Jan. 18–20, 2017). ACM, 300–315; https://doi.org/10.1145/3009837
7. Clune, T. and Rood, R. Software testing and verification in climate model development. *IEEE Software* 28, 6 (2011), 49–55; https://doi.org/10.1109/MS.2011.117
8. Collberg, C.A. and Proebsting, T.A. Repeatability in computer systems research. *Commun. ACM* 59, 3 (Mar. 2016), 62–69; https://doi.org/10.1145/2812803
9. Compcert. The Compcert Project, 2019; <http://www.compcert.inria.fr>
10. Easterbrook, S.M., Edwards, P.N., V. Balaji, V. and R. Budich, R. Climate change: Science and software. *IEEE Software* 28, 6 (2011), 32–35.
11. Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J., and Taylor, K.E. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* 9, 5 (2016), 1937–1958; https://doi.org/10.5194/gmd-9-1937-2016.
12. fpanalysistools.org. Tutorial on Floating-Point Analysis Tools; <http://fpanalysistools.org/>
13. Hurrell, J. et al. The Community Earth System Model: A framework for collaborative research. *Bulletin of the American Meteorological Society* 94 (2013), 1339–1360; https://doi.org/10.1175/BAMS-D-12-00121.1
14. Intel. BFloat16—Hardware Numerics. White Paper, Document Number: 338302-001US, Revision 1.0, 2018; https://intel.ly/36IJ37r.
15. IPCC 2019. Intergovernmental Panel on Climate Change; <http://www.ipcc.ch/about>.
16. James, D. et al. Standing Together for Reproducibility in Large-Scale Computing: Report on reproducibility@XSEDE. CoRR abs/1412.5557 (2014), 16. arXiv:1412.5557; <http://arxiv.org/abs/1412.5557>.
17. Kahan, W. Lecture notes on the status of IEEE Standard 754 for binary floating-point arithmetic.1997; https://people.eecs.berkeley.edu/~wkahan/ieee754status/IEEE754.PDF
18. Kahan, W. How futile are mindless assessments of roundoff in floating-point computation? 2006; https://people.eecs.berkeley.edu/~wkahan/Mindless.pdf

19. Kim, Y et al. KGEN: A Python tool for automated Fortran kernel generation and verification. In *Proceedings of the 2016 Intern. Conf. Computational Science* 80, 1450–1460; https://doi.org/10.1016/j.procs.2016.05.466.
20. Menon, H., et al. ADAPT: Algorithmic Differentiation Applied to Floating-point Precision Tuning. In *Proceedings of the 2018 Intern. Conf. High Performance Computing, Networking, Storage, and Analysis*, Article 48. IEEE Press, Piscataway, NJ, USA; https://doi.org/10.1109/SC.2018.00051.
21. Milroy, D.J., Baker, A.H., Hammerling, D.M. and Jessup, E.R. Nine time steps: Ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0). *Geoscientific Model Development* 11, 2 (2018), 697–711; https://doi.org/10.5194/gmd-11-697-2018.
22. Milroy, D.J., Baker, A.H., Hammerling, D.M., Kim, Y., Jessup, E.R., and Hauser, T. Making root cause analysis feasible for large code bases: A solution approach for a climate model. In *Proceedings of the 28th Intern. Symp. High-Performance Parallel and Distributed Computing*. ACM, 2019, 73–84; https://doi.org/10.1145/3307681.3325399.
23. Pipitone, J. and Easterbrook, S. Assessing climate model software quality: A defect density analysis of three models. *Geoscientific Model Development* 5, 4 (2012), 1009–1022; https://doi.org/10.5194/gmd-5-1009-2012.
24. PRUNERS. FLIT and ReMPI Projects page, 2019; https://pruners.github.io/flit/
25. Biere, A., Huelde, M., van Maaren, H. and Walsh, T. *Handbook of Satisfiability*. IOS Press, 2008.
26. Riedy, E.J. and Demmel, J. Augmented arithmetic operations proposed for IEEE-754 2018. In *Proceedings of the 25th IEEE Symp. Computer Arithmetic* (Amherst, MA, USA, June 25–27, 2018), 45–52; https://doi.org/10.1109/ARITH.2018.8464813.
27. Rubio-González, C. et al. Precimonious: Tuning assistant for floating-point precision. In *Proceedings of the Intern. Conf. High Performance Computing, Networking, Storage and Analysis* (Denver, CO, USA, Nov. 17–21, 2013). W. Grop and S.S. Matsuoka (Eds.), ACM, 27:1–27:12; https://doi.org/10.1145/2503210.2503296.
28. Small, R.J. et al. A new synoptic scale resolving global climate simulation using the Community Earth System Model. *J. Advances in Modeling Earth Systems* 6, 4 (2014), 1065–1094; https://doi.org/10.1002/2014MS000363.
29. Solovoyev, A., Baranowski, M.S., Briggs, I., Jacobsen, C., Rakamaric, Z. and Gopalakrishnan, G. Rigorous estimation of floating-point round-off errors with Symbolic Taylor Expansions. *ACM Trans. Program. Lang. Syst.* 41, 1, Article 2 (Dec. 2018); https://doi.org/10.1145/3230733.
30. Zeller, A. and Hildebrandt, R. Simplifying and isolating failure-inducing input. *IEEE Trans. Software Engineering* 28, 2 (2002), 183–200.

**Dong H. Ahn** is a computer scientist at the Lawrence Livermore National Laboratory, Livermore, CA, USA.

**Allison Baker** is Project Scientist III in the Computational Information Systems Laboratory, National Center for Atmospheric Research, Boulder, CO, USA.

**Michael Bentley** is pursuing a Ph.D. at the School of Computing, University of Utah, Salt Lake City, UT, USA.

**Ian Briggs** is pursuing a Ph.D. at the School of Computing, University of Utah, UT, USA.

**Ganesh Gopalakrishnan** is a professor at the School of Computing, University of Utah, Salt Lake City, UT, USA.

**Dorit Hammerling** is an associate professor in the Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO, USA.

**Ignacio Laguna** is a computer scientist at the Lawrence Livermore National Laboratory, Livermore, CA, USA.

**Gregory L. Lee** is a computer scientist at the Lawrence Livermore National Laboratory, Livermore, CA, USA.

**Daniel Milroy** is a postdoctoral researcher at the Lawrence Livermore National Laboratory, Livermore, CA, USA.

**Mariana Vertenstein** leads the CESM Software Engineering Group (since 2004) in the Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO, USA.

# 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

## Tracing the triumphs and challenges of two decades of Semantic Web research and applications.

BY PASCAL HITZLER

# A Review of the Semantic Web Field

LET US BEGIN this review by defining the subject matter. The term *Semantic Web* as used in this article is a field of research rather than a concrete artifact—in a similar way, say, *Artificial Intelligence* denotes a field of research rather than a concrete artifact. A concrete artifact, which may deserve to be called “The Semantic Web” may or may not come into existence someday, and indeed some members of the research field may argue that part of it has already been built. Sometimes the term *Semantic Web technologies* is used to describe the set of methods and tools arising out of the field in an attempt to avoid terminological confusion. We will come back to all this in the article in some way; however, the focus here is to review the research field.

This review will be rather subjective, as the field is very diverse not only in methods and goals being researched and applied, but also because

the field is home to a large number of different but interconnected subcommunities, each of which would probably produce a rather different narrative of the history and the current state of the art of the field. I therefore do not strive to achieve the impossible task of presenting something close to a consensus—such a thing still seems elusive. However, I do point out here, and sometimes within the narrative, that there are a good number of alternative perspectives.

The review is also very selective, because Semantic Web is a rich field of diverse research and applications, borrowing from many disciplines within or adjacent to computer science. In a brief review like this one cannot possibly be exhaustive or give due credit to all important individual contributions. I do hope I have captured what many would consider key areas of the Semantic Web field. For the reader interested in obtaining a more detailed overview, I recommend perusing the major publication outlets in the field: The *Semantic Web* journal,<sup>a</sup> the *Journal of Web Semantics*,<sup>b</sup> and the proceedings of the annual International Semantic Web Conference.<sup>c</sup> This is by no means an exhaustive list, but I believe it to be uncontroversial that these are the most central publication venues for the field.

Now that we understand that Semantic Web is a field of research, what is it about? Answers to this question are again necessarily subjective as there is no clear consensus on this in the field.<sup>d</sup>

One perspective is that the field is all about the long-term goal of creating The Semantic Web (as an artifact) together with all the necessary tools and methods

a <http://www.semantic-web-journal.net/>

b <https://www.journals.elsevier.com/journal-of-web-semantics>

c <http://swsa.semanticweb.org/content/international-semantic-web-conference-iswc>

d I would like to emphasize this lack of consensus is as much a boon for the field, giving it diversity, as it is sometimes a disadvantage.



Knowledge  
Graphs

Semantic  
Web

Linked  
Data

Ontologies

required for creation, maintenance, and application. In this particular narrative, The Semantic Web is usually envisioned as an enhancement of the current World Wide Web with machine-understandable information (as opposed to most of the current Web, which is mostly targeted at human consumption), together with services—intelligent agents—utilizing this information. This perspective can be traced back to a 2001 *Scientific American* article,<sup>1</sup> which arguably marks the birth of the field. Provision of machine understandable information in this case is done by endowing data with expressive metadata for the data. In the Semantic Web, this metadata is generally in the form of ontologies, or at least a formal language with a logic-based semantics that admits reasoning over the meaning of the data. (Formal metadata is discussed later.) This, together with the understanding that intelligent agents would utilize the information, perceives the Semantic Web field as having a significant overlap with the field of Artificial Intelligence. Indeed, for most of the major artificial intelligence conferences held in the last 20 years ran explicit “Semantic Web” tracks.

An alternative and perhaps more recent perspective on the question of what the field is about rests on the observation that the methods and tools developed by the field have applications not tied to the World Wide Web, and which also can provide added value even without having to establish intelligent agents utilizing machine-understandable data. Indeed, early industry interest in the field, which was substantial from the very outset, was aimed at applying Semantic Web technologies to information integration and management. From this perspective, one could argue the field is about establishing efficient (that is, low cost) methods and tools for data sharing, discovery, integration, and reuse, and the World Wide Web may or may not be a data transmission vehicle in this context. This understanding of the field moves it closer to databases, or the data management part of data science.

A much more restrictive, but perhaps practically rather astute, delineation of the field may be made by characterizing it as investigating foundations and applications of ontologies, linked data, and knowledge graphs (all

discussed later), with the W3C standards<sup>e</sup> RDF, OWL, and SPARQL at its core.

Perhaps, each of these three perspectives has merit, and the field exists in a confluence of these, with ontologies, linked data, knowledge graphs, being key concepts for the field, W3C standards around RDF, OWL, and SPARQL constituting technical exchange formats that unify the field on a syntactic (and to a certain extent, semantic) level; the application purpose of the field is in establishing efficient methods for data sharing, discovery, integration, and reuse (whether for the Web or not); and a long-term vision that serves as a driver is the establishing of The Semantic Web as an artifact complete with intelligent agent applications at some point in the (perhaps, distant) future.

In the rest of this article, I will lay out a timeline of the field’s history, covering key concepts, standards, and prominent outcomes. I will also discuss some selected application areas as well as the road and challenges that lie ahead.

### A Subjective Timeline

Declaring any specific point in time as the birth of a field of research is of course debatable at best. Nevertheless, a 2001 *Scientific American* article by Berners-Lee et al.<sup>1</sup> is an early landmark and has provided significant visibility for the nascent field. And, yes, it was around the early 2000s when the field was in a very substantial initial upswing in terms of community size, academic productivity, and initial industry interest.

But there were earlier efforts. The DARPA Agent Markup Language (DAML) program<sup>f</sup> ran from 2000 to 2006 with the declared goal of developing a Semantic Web language and corresponding tools. The European Union-funded On-To-Knowledge project,<sup>g</sup> running from 2000–2002, gave rise to the OIL language that was later merged with DAML, eventually giving rise to the Web Ontology Language (OWL) W3C standard. The more general idea of endowing data on the Web with machine-readable or “understandable” metadata can be traced back to the beginnings of the World

Wide Web itself. For example, a first draft of the Resource Description Framework (RDF) was published as early as 1997.<sup>h</sup>

Our story of the field will commence from the early 2000s, and we group the narrative into three overlapping phases, each driven by a key concept; that is, under this reconstruction, the field has shifted its main focus at least twice. From this perspective, the first phase was driven by *ontologies* and it spans the early to mid 2000s; the second phase was driven by *linked data* and stretches into the early 2010s. The third phase was and is still driven by *knowledge graphs*.

**Ontologies.** For most of the 2000s, work in the field had the notion of ontology at its center, which, of course, has much older roots. According to a many-cited source from 1993,<sup>5</sup> an ontology is a formal, explicit specification of a shared conceptualization—though one may argue that this definition still needs interpretation and is rather generic. In a more precise sense (and perhaps a bit post-hoc), an ontology is really a knowledge base (in the sense of symbolic artificial intelligence) of concepts (that is, types or classes, such as “mammal” and “live birth”) and their relationships (such as, “mammals give live birth”), specified in a knowledge representation language based on a formal logic. In a Semantic Web context, ontologies are a main vehicle for data integration, sharing, and discovery, and a driving idea is that ontologies themselves should be reusable by others.

In 2004, the Web Ontology Language OWL became a W3C standard (the revision OWL 2<sup>11</sup> was established in 2012), providing further fuel for the field. OWL in its core is based on a *description logic*, that is, on a sub-language of first-order predicate logic<sup>i</sup> using only unary and binary predicates and a restricted use of quantifiers, designed in such a way that logical deductive reasoning over the language is decidable.<sup>12</sup> Even after the standard was established, the community continued to have discussions whether description logics were the best paradigm choice, with rule-based languages being a major contender.<sup>28</sup> The discussion

e The World Wide Web Consortium (W3C) calls its standards “Recommendations.”

f <http://www.daml.org/>

g <https://cordis.europa.eu/project/id/IST-1999-10132>

h <https://www.w3.org/TR/WD-rdf-syntax-971002/>

i With some mild extensions not found in standard first-order predicate logic, such as counting quantifiers.

eventually settled, but the Rule Interchange Format RIF,<sup>25</sup> which was later established as a rule-based W3C standard gained relatively little traction.<sup>j</sup>

Also in 2004, the Resource Description Framework (RDF) became a W3C standard (the revision RDF 1.1<sup>32</sup> was completed in 2014). In essence, RDF is a syntax for expressing directed, labeled, and typed graphs.<sup>k</sup> RDF is more or less<sup>l</sup> compatible with OWL, by using OWL to specify an ontology of types and their relationships, and by then using these types as types in the RDF graph, and the relationships as edges. From this perspective, an OWL ontology can serve as a *schema* (or a logic of types) for the RDF (typed) graph.<sup>m</sup>

A W3C standard for an RDF query language, called SPARQL, followed in 2008 (with an update in 2013,<sup>36</sup> which then also became more fully compatible with OWL). Additional standards in the vicinity of RDF, OWL, and SPARQL have been, or are being, developed, some of which have gained significant traction, for example, ontologies such as the Semantic Sensor Networks ontology<sup>7</sup> or the Provenance ontology,<sup>20</sup> or the SKOS Simple Knowledge Organization System.<sup>24</sup>

With all these key standards developed under the W3C, basic compatibility between them and other key W3C standards has been maintained. For example, XML serves as a syntactic serialization and interchange format for RDF

j Evidence, for example, is given by comparing Google Scholar citation counts for the standards documents, which are two orders of magnitude lower for RIF.

k The full standard is more complicated; for example, it allows things like using edge labels, or node types, also as nodes from which other edges originate, which would be in violation of what is usually considered a graph. Excessive use of such departures from standard graph structures are usually used sparingly, as the results are often hard to interpret.

l Syntactically, they are fully compatible, as RDF is a syntactic serialization format for OWL. However, RDF and OWL each carry a (more precisely, several) formal semantics that are not fully compatible between the languages. To the best of my knowledge, there is no single reference which discusses the exact relationship in detail, but Hitzler et al.<sup>12</sup> gives some indications.

m RDF Schema,<sup>32</sup> which is part of the RDF standard, can serve this purpose as well but is much less expressive than OWL, and in terms of semantics not fully compatible with it – see the previous footnote.



**In a Semantic Web context, ontologies are a main vehicle for data integration, sharing, and discovery, and a driving idea is that ontologies themselves should be reusable by others.**



and OWL. All W3C Semantic Web standards also use IRIs as identifiers for labels in an RDF graph, for OWL class names, for datatype identifiers among others.

The DARPA DAML program ended in 2006, and subsequently there were few if any large-scale funding lines for fundamental Semantic Web research in the U.S. As a consequence, much of the corresponding research in the U.S. moved either to application areas such as data management in healthcare or defense, or into adjacent fields altogether. In contrast, the European Union Framework Programmes, in particular FP 6 (2002–2006) and FP 7 (2007–2013), provided significant funding for both foundational and application-oriented Semantic Web research. One of the results of this divergence in funding priorities is still mirrored in the composition of the Semantic Web research community, which is predominantly European. The size of the community is difficult to assess, but since the mid-2000s, the field's key conference—the International Semantic Web Conference—has drawn over 600 participants on average each year.<sup>n</sup> Given the interdisciplinary nature and diverse applications of the field, it is to be noted that much Semantic Web research or applications are published in venues for adjacent research or application fields.

Industry interest has been significant from the outset, but it is next to impossible to reconstruct reliable data on the precise level of related industry activity. University spin-offs applied state-of-the-art research from the outset, and graduating Ph.D. students—in particular, the significant number produced in Europe—were finding corresponding industry jobs. Major and smaller companies have been involved in large-scale foundational or applied research projects, in particular under EU FP 6 and 7. Industry interest has changed focus with the research community, and we will come back to this throughout the narrative.

Some large-scale ontologies, often with roots predating the Semantic Web community, matured during this time. For example, the Gene Ontology<sup>35</sup> had

n The much newer annual China Conference on Knowledge Graph and Semantic Computing, established in 2013, with primarily national focus, has by now grown to almost 1,500 participants.



its beginnings in 1998 and is now a very prominent resource. Another example is SNOMED CT,<sup>o</sup> which can be traced back to the 1960s but is now fully formalized in OWL and widely used for electronic health records.<sup>33</sup>

As is so often the case in computer science research, initial over-hyped expectations on short-term massive breakthrough results gave way, around the mid-2000s, to a more sober perspective. Ontologies in the form that were mostly developed during this time—meaning often based on ad-hoc modeling as methodologies for their development were researched but had not yet led to tangible results—turned out to be difficult to maintain and re-use. This, combined with the considerable up-front cost at that time to develop good ontologies,<sup>p</sup> paved the way for a shift in attention by the research community, which can be understood as perhaps antithetical to the strongly ontology-based approach of the early 2000s.

**Linked Data.** The year 2006 saw the birth of “linked data” (or “linked open data” if the emphasis is on open, public, availability under free licenses). Linked data<sup>3</sup> would soon become a major driver for Semantic Web research and applications and persist as such until the early 2010s.

What is usually associated with the term “linked data” is that linked data consists of a (by now rather large) set of RDF graphs that are linked in the sense that many IRI identifiers in the graphs

also appear also in other, sometimes multiple, graphs. In a sense, the collection of all these linked RDF graphs can be understood as one very big RDF graph.

The number of publicly available linked RDF graphs has been showing significant growth in particular during the first decade as shown in Figure 1; the data is from the Linked Open Data Cloud website,<sup>q</sup> which does not account for all RDF datasets on the Web. A 2015 paper<sup>29</sup> reports on “more than 37 billion triples<sup>r</sup> from over 650,000 data documents,” which is also only a selection of all RDF graph triples that can be freely accessed on the World Wide Web. Large data providers, for example, often provide only a query interface based on SPARQL (a “SPARQL endpoint”) or use RDF for internal data organization but provide it to the outside only via human-readable Web pages. Datasets in the Linked Open Data Cloud cover a wide variety of topics, including geography, government, life sciences, linguistics, media, scientific publications, and social networking.

One of the most well-known and used linked datasets is DBpedia,<sup>22</sup> which is a linked dataset extracted from Wikipedia (and, more recently, also Wikidata). The April 2016 release<sup>s</sup> covers about six million entities and about 9.5 billion RDF triples. Due to its extensive topic coverage (essentially, everything in Wikipedia) and the fact it was one of the very first linked datasets to be made available, DBpedia plays a

q <https://lod-cloud.net/>

r In RDF terminology, a triple consists of a node-edge-node piece of an RDF graph.

s <https://blog.dbpedia.org/2016/10/19/yeah-we-did-it-again-new-2016-04-dbpdiarelease/>

central role in the Linked Open Data Cloud of interlinked datasets: Many other datasets link to it so that it has become a kind of hub for linked data.

There was significant industry interest in linked data from the outset. For example, BBC<sup>t</sup> was one of the first significant industry contributors to the Linked Data Cloud and the New York Times Company<sup>31</sup> and Facebook<sup>40</sup> were early adopters. However, industry interest seemed mostly be about utilizing linked data *technology* for data integration and management, often without it being visible on the open World Wide Web.

During the Linked Data era, ontologies played a much less prominent role. They often were used as schemas in that they informed the internal structure of RDF datasets, however, the information in RDF graphs in the Linked Data Cloud was shallow and relatively simplistic compared to the overpromises and depth of research from the Ontologies era. The credo sometimes voiced during this time was that ontologies cannot be reused, and that a much simpler approach based mainly on utilizing RDF and links between datasets held much more realistic promises for data integration, management, and applications on and off the Web. It was also during this time that RDF-based data organization vocabularies with little relation to ontologies, such as SKOS,<sup>24</sup> were developed.

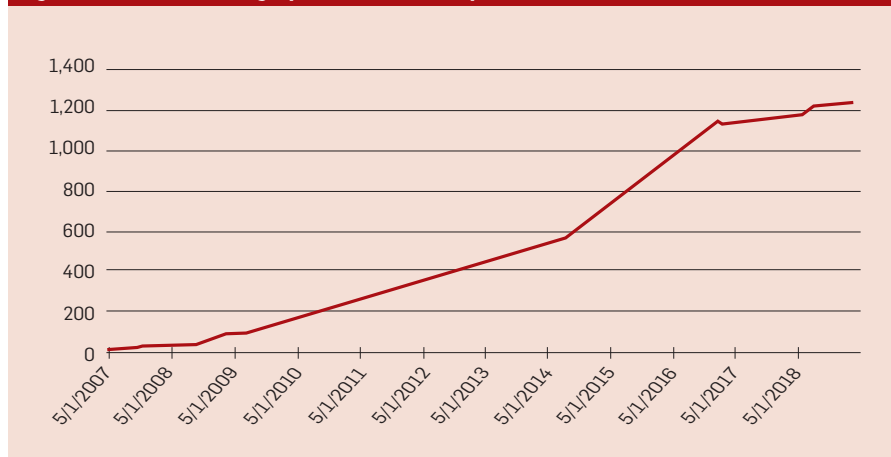
It was also during this time (2011) when schema.org appeared on the scene.<sup>6</sup> Initially driven by Bing, Google, and Yahoo!—and slightly later joined by Yandex—schema.org made public a relatively simple ontology<sup>u</sup> and suggested that website providers annotate (that is, link) entities on their sites with the schema.org vocabulary. In return, the Web search engine providers behind schema.org promised to improve search results by utilizing the annotations as metadata. Schema.org saw considerable initial uptake: In 2015, Guha et al.<sup>6</sup> reported over 30% of pages have schema.org annotations.

Another prominent effort launched in 2012 is Wikidata,<sup>39</sup> which started as a

t <https://www.bbc.co.uk/academy/en/articles/art20130724121658626>

u As of the writing of this article it has 614 classes and 902 relations and consists primarily of a type hierarchy.

**Figure 1. Number of RDF graphs in the Linked Open Data Cloud over time.**



project at Wikimedia Deutschland funded among others by Google, Yandex, and the Allen Institute for AI. Wikidata is based on a similar idea as Wikipedia, namely, to crowdsource information. However, while Wikipedia is providing encyclopedia-style texts (with human readers as the main consumers), Wikidata is about creating structured data that can be used by programs or in other projects. For example, many other Wikimedia efforts, including Wikipedia, use Wikidata to provide some of the information they present to human readers. As of the time of this writing, Wikidata has over 66 million data items, has had over one billion edits since project launch, and has over 20,000 active users.<sup>v</sup> Database downloads are available in several W3C standards, including RDF.

During the early 2010s, the initial hype about linked data began to give way to a more sober perspective. While there were indeed some prominent uses and applications of linked data, it still turned out that integrating and utilizing it took more effort than some initially expected. Arguably, shallow non-expressive schemas often used for linked data appeared to be a major obstacle to reusability,<sup>16</sup> and initial hopes that interlinks between datasets would somehow account for this weakness did not really seem to materialize. This observation should not be understood as demeaning the significant advances linked data has brought to the field and its applications: Just having data available in some structured format that follows a prominent standard means it can be accessed, integrated, and curated with available tools, and then made use of—and this is much easier than if data is provided in syntactically and conceptually much more heterogeneous form. But the quest for more efficient approaches to data sharing, discovery, integration, and reuse was of course as important as ever, and is commencing.

**Knowledge Graphs.** In 2012, a new term appeared on the scene when Google launched its Knowledge Graph. Pieces of the Google Knowledge Graph can be seen, for example, by searching for prominent entities on [google.com](http://google.com): Next to the search results linking to Web pages a so-called infobox is displayed that shows information from

the Google Knowledge Graph. An example of such an infobox is given in Figure 2—this was retrieved by searching the term Kofi Annan. One can navigate from this node to other nodes in the graph by following one of the active hyperlinks, for example, to Nane Maria Annan who is listed with a spouse relationship to the Kofi Annan node. After following this link, a new infobox for Nane Maria Annan is displayed next to the usual search results for the same term.

While Google does not provide the Knowledge Graph for download, it does provide an API to access content<sup>w</sup>—the API uses standard [schema.org](http://schema.org) types and is compliant with JSON-LD,<sup>34</sup> which is essentially an alternative syntax for RDF standardized by the W3C.

Knowledge graph technology has found a prominent place in industry, including leading information technology companies other than Google, such as Microsoft, IBM, Facebook, and eBay.<sup>27</sup> However, given the history of Semantic Web technologies, and in particular of linked data and ontologies discussed earlier, it seems that knowledge graph is mostly a new framing of ideas coming directly out of the Semantic Web field,<sup>x</sup> with some notable shifts in emphasis.

One of the differences is about openness: As the term *Linked Open Data* has suggested from the very beginning, the linked data efforts by the Semantic Web community mostly had open sharing of data for reuse as one of its goals, which means that linked data is mostly made freely available for download or by SPARQL endpoint, and the use of non-restricting licenses is considered of importance in the community. Wikidata as a knowledge graph is also unowned, and open. In contrast, the more recent activities around knowledge graphs are often industry-led, and the prime showcases are not really open in this sense.<sup>27</sup>

Another difference is one of central control versus bottom-up community contributions: The Linked Data Cloud is in a sense the currently largest existing knowledge graph known, but it is hardly a concise entity. Rather, it consists of

**Figure 2. Google Knowledge Graph node as shown after searching on [google.com](http://google.com) for the term “Kofi Annan.”**



loosely interlinked individual subgraphs, each of which is governed by its very own structure, representation schema, and so on. Knowledge graphs, in contrast, are usually understood to be much more internally consistent, and more tightly controlled, artifacts. As a consequence, the value of external links—that is, to external graphs without tight quality control—is put into doubt,<sup>y</sup> while quality of content and/or the underlying schema comes more into focus.

The biggest difference is probably the transition from academic research (which mostly drove the linked data effort) to use in industry. As such, recent activities around knowledge graphs are fueled by the strong industrial use cases and their demonstrated or perceived added value, even though there is, to the best of my knowledge, no published formal evaluation of their benefits.

Yet many of the challenges and issues concerning knowledge graphs remain the same as they were for linked data; for example, all items on the list of current challenges listed in Noy et al.<sup>27</sup> are very well-known in the Semantic Web field, many with substantial bodies of research having been undertaken.

### Selected Relationships to other Fields and Disciplines

As we discussed, the Semantic Web field is not primarily driven by certain methods inherent to the field, which distinguishes it from some other areas such

<sup>w</sup> <https://developers.google.com/knowledge-graph>

<sup>x</sup> The term *knowledge graph* is of course also not new as such, it was already used, for example, in the 1980s with a similar general meaning.

<sup>y</sup> Early indicators of this have shown for example that many of the same-as links contained in the Linked Data Cloud link entities which should not as such be considered exactly the same.<sup>8</sup>

<sup>v</sup> <https://www.wikidata.org/wiki/Wikidata:Statistics>

as machine learning. Rather, it is driven by a shared vision,<sup>z</sup> and as such it borrows from other disciplines as needed.<sup>aa</sup>

For example, the Semantic Web field has strong relations to knowledge representation and reasoning as a sub-discipline of artificial intelligence, as knowledge graph and ontology representation languages can be understood—and are closely related to—knowledge representation languages, with description logics, as the logics underpinning the Web Ontology Language OWL, playing a central role. Semantic Web application requirements have also driven or inspired description logic research, as well as investigations into bridging between different knowledge representation approaches such as rules and description logics.<sup>19</sup>

The field of databases is clearly closely related, where topics such as (meta)data management and graph-structured data have a natural home but are also of importance for the Semantic Web field. However, the emphasis in Semantic Web research is strongly focused on conceptual integration of heterogeneous sources; for example, how to overcome different ways to organize data; in Big Data terminology, Semantic Web emphasis is primarily on the variety aspect of data.<sup>17</sup>

Natural language processing as an application tool plays an important role, for example, for knowledge graph and ontology integration, for natural language query answering, as well as for automated knowledge graph or ontology construction from texts.

Machine learning, and in particular deep learning, are being investigated as to their capability to improve hard tasks arriving in a Semantic Web context, such as knowledge graph completion (in the sense of adding missing relations), dealing with noisy data, and so on.<sup>4,10</sup> At the same time, Semantic Web technologies are being investigated as to their potential to advance explainable AI.<sup>10,21</sup>

Some aspects of cyber-physical systems and the Internet of Things are

being researched on using Semantic Web technologies, for example, in the context of smart manufacturing (Industry 4.0), smart energy grids, and building management.<sup>30</sup>

Some areas in the life sciences have already a considerable history of benefiting from Semantic Web technologies, for example, the previously noted SNOMED-CT and Gene Ontology. Generally speaking, biomedical fields were early adopters of Semantic Web concepts. Another prominent example would be the development of the ICD11, which was driven by Semantic Web technologies.<sup>38</sup>

Other current or potential application areas for Semantic Web technologies can be found wherever there is a need for data sharing, discovery, integration, and reuse, for example, in geosciences or in digital humanities.<sup>15</sup>

### Some of the Road Ahead

Undoubtedly, the grand goal of the Semantic Web field—be it the creation of The Semantic Web as an artifact, or providing solutions for data sharing, discovery, integration, and reuse, which make it completely easy and painless—has not yet been achieved. This does not mean that intermediate results are not of practical use or even industrial value, as the discussions about knowledge graphs, schema.org, and the life science ontologies demonstrate.

Yet, to advance toward the larger goals, further advances are required in virtually every subfield Semantic Web. For many of these, discussions of some of the most pressing challenges can be found, for example, in Bernstein et al.<sup>2</sup> in the contributions to the January 2020 special issue of the *Semantic Web* journal<sup>ab</sup> or in Noy et al.<sup>27</sup> for industrial knowledge graphs, in Thieblin et al.<sup>37</sup> for ontology alignment, in Martinez-Rodriguez et al.<sup>23</sup> for information extraction, in Höffner et al.<sup>13</sup> for question answering, or in Hammer et al.<sup>9</sup> for ontology design patterns and more. Rather than to repeat or recompile these lists, let us focus on the challenge that I personally consider to be the current, short-term, major roadblock for the field at large.

There is a wealth of knowledge—hard and soft—in the Semantic Web

community and its application communities about how to approach issues around efficient data management. Yet, new adopters often find themselves confronted with a cacophony of voices pitching different approaches, little guidance as to the pros and cons of these different approaches, and a bag of tools which range from crude unfit-for-practice research prototypes to well-designed software for particular subproblems, but again with little guidance which tools, and which approaches, will help them best in achieving their particular goals.

Thus, what the Semantic Web field most needs, at this stage, is consolidation. And as an inherently application-driven field, this consolidation will have to happen across its subfields, resulting in application-oriented processes that are well-documented as to their goals and pros and cons, and which are accompanied by easy-to-use and well-integrated tools supporting the whole process. For example, some of the prominent and popular software available, such as the Protégé ontology editor,<sup>26</sup> the OWL API,<sup>14</sup> Wikibase, which is the engine underlying Wikidata,<sup>ac</sup> or the ELK reasoner,<sup>18</sup> are powerful and extremely helpful, but fall far short from working easily with each other in some cases, even though they all use RDF and OWL for serializations.

Who could be the drivers of such consolidation? For academics, there is often limited incentive to develop and maintain stable, easy-to-use software, as academic credit—mostly measured in publications and in the sum of acquired external funding—often does not align well with these activities. Likewise, complex processes are inherently difficult to evaluate, which means that top-tier publication options for such kinds of work are limited. Writing high-quality introductory textbooks as a means to consolidate a field is very time-consuming and returns very little academic credit. Yet, the academic community does provide a basis for consolidation, by developing solutions that bridge between paradigms, and by partnering with application areas to develop and materialize use-cases.

Consolidation of sorts is already

z Another discipline not primarily driven by methods, but rather by shared vision or goals is, cybersecurity.

aa For example, see the ISWC 2006 keynote by Rudi Studer on Semantic Web: Customers and Suppliers, see [http://videlectures.net/iswc06\\_studer\\_sc/](http://videlectures.net/iswc06_studer_sc/).

ab <http://www.semantic-web-journal.net/issues>

ac <https://wikiba.se/>



happening in industry, as witnessed by the adoption of Semantic Web technologies in start-ups and multinationals. Technical details, not even to speak of in-house software, underlying this adoption, for example, as in the case of the industrial knowledge graphs discussed in Noy et al.,<sup>27</sup> are however usually not shared, presumably to protect the own competitive edge. If this is indeed the case, then it may only be a matter of time before corresponding software solutions become more widely available.

## Conclusion


Within its first approximate 20 years of existence, the Semantic Web field has produced a wealth of knowledge regarding efficient data management for data sharing, discovery, integration, and reuse. The contributions of the field are best understood by means of the applications they have given rise to, including Schema.org, industrial knowledge graphs, Wikidata, ontology modeling applications, among other fields discussed throughout this article.

It is natural to also ask about the key scientific discoveries that have provided the foundations for these applications; however, this question is much more difficult to answer. What I hope has become clear from the narrative, advances in the pursuit of the Semantic Web theme require contributions from many computer science subfields, and one of the key quests is about finding out how to piece together contributions, or modifications thereof, in order to provide applicable solutions. In this sense, the applications (including those mentioned herein) showcase the major scientific progress of the field as a whole.

Of course, many of the contributing fields have individually made major advances in the past 20 years, and sometimes central individual publications have decisively shaped the narrative of a subfield. Reporting in more detail on such advances would be a worthwhile endeavor but constitute a separate piece in its own right. The interested reader is encouraged to follow up on the references given, which in turn will point to the key individual technological contributions that lead to the existing widely used standards, the landmark applications reported here-

in, and the current discussion on open technical issues in the field to which references have been included.

The field is seeing mainstream industrial adoption, as laid out in the narrative. However, the quest for more efficient data management solutions is far from over and continues to be a driver for the field.

**Acknowledgment.** This work was supported by the National Science Foundation under award OIA-2033521. 

## References

- Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American* 284, 5 (May 2001), 34–43.
- Bernstein, A., Hendler, J., and Noy, N. A new look at the Semantic Web. *Commun. ACM* 59, 9 (Sept. 2016), 35–37.
- Bizer, C., Heath, T., and Berners-Lee, T. Linked Data—The story so far. *Int. J. Semantic Web Inf. Syst.*, 3 (2009), 1–22.
- d'Amato, C. 2020. Machine learning for the Semantic Web: Lessons learnt and next research directions. *Semantic Web* 11, 1 (2020), 195–203.
- Gruber, T. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5, 2 (1993), 199–220.
- Guha, R., Brickley, D., and Macbeth, S. 2016. Schema.org: evolution of structured data on the web. *Commun. ACM* 59, 2 (2016), 44–51. <https://doi.org/10.1145/2844544>
- Haller, A., Janowicz, K., Cox, S., Phuoc, D., Taylor, K., and Lefrancois, M (Eds.). 2017. *Semantic Sensor Network Ontology*. W3C Recommendation 19 October 2017. Available from <http://www.w3.org/TR/vocabssn/>.
- Halpin, H., Hayes, P., and Thompson, H. When owl: Same as isn't the same redux: A preliminary theory of identity and inference on the Semantic Web. In *Proceedings of Workshop on Discovering Meaning on the Go in Large Heterogeneous Data*. (Barcelona, Spain, July 16, 2011), 25–30.
- Hammar, K. et al. Collected research questions concerning ontology design patterns. *Ontology Engineering with Ontology Design Patterns—Foundations and Applications*. P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi, and V. Presutti (Eds.). *Studies on the Semantic Web* 25. IOS Press, 2016, 189–198.
- Hitzler, P., Bianchi, F., Ebrahimi, M., and Sarker, M. Neural-symbolic integration and the Semantic Web. *Semantic Web* 11, 1 (2020), 3–11.
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P., and Rudolph, S. (Eds.). *OWL 2 Web Ontology Language: Primer (2nd Ed.)*. W3C Recommendation 11 (Dec. 2012); <http://www.w3.org/TR/owl2-primer/>.
- Hitzler, P., Krötzsch, M., and Rudolph, S. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2010.
- Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., and Ngomo, A. Survey on challenges of question answering in the Semantic Web. *Semantic Web* 8, 6 (2017), 895–920.
- Horrige, M. and Bechhofer, S. The OWL API: A Java API for OWL ontologies. *Semantic Web* 2, 1 (2011), 11–21.
- Hyyönen, E. Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web* 11, 1 (2020), 187–193.
- Jain, P., Hitzler, P., Yeh, P., Verma, K., and Sheth, A. Linked Data Is Merely More Data. Papers from the 2010 AAAI Spring Symposium, Technical Report SS-10-07. *Linked Data Meets Artificial Intelligence*. (Stanford, CA, USA, Mar. 22–24, 2010). AAAI.
- Janowicz, K., van Harmelen, F., Hendler, J., and Hitzler, P. Why the data train needs semantic rails. *AI Magazine* 36, 1 (2015), 5–14.
- Kazakov, Y., Krötzsch, M., and Simancik, F. The incredible ELK—From polynomial procedures to efficient reasoning with EL ontologies. *J. Autom. Reasoning* 53, 1 (2014), 1–61.
- Krisnadhi, A., Maier, F., and Hitzler, P. OWL and rules. In *Proceedings of the 7th Intern. Summer School: Reasoning Web Semantic Technologies for the Web of Data*. (Galway, Ireland, Aug. 23–27, 2011). A. Polleres, C. d'Amato, M. Arenas, S. Handschuh, P. Kroner, S. Ossowski, and P.F. Patel-Schneider (Eds.). LNCS 6848. Springer, 382–415.
- Lebo, T., Sahoo, S., and McGuinness, D. (Eds.). *PROV-O: The PROV Ontology*. W3C Recommendation (Apr. 30, 2013); <http://www.w3.org/TR/prov-o/>.
- Lecue, F. On the role of knowledge graphs in explainable AI. *Semantic Web* 11, 1 (2020), 41–51.
- Lehmann, J. et al. DBpedia—A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.
- Martinez-Rodriguez, J., Hogan, A., and Lopez-Arevalo, I. Information extraction meets the Semantic Web: A Survey. *Semantic Web* 11, 2 (2020), 255–335.
- Miles, A. and Bechhofer, S. (Eds.). SKOS Simple Knowledge Organization System. W3C Recommendation (Aug. 18, 2009); <http://www.w3.org/TR/skos-reference/>.
- Morgenstern, L., Welty, C., Boley, H., and Hallmark, G. (Eds.). *RIF Primer (2nd Ed.)*. W3C Working Group Note 5 (Feb. 2013); <http://www.w3.org/TR/rif-primer/>.
- Musen, M. The Protégé project: a look back and a look forward. *AI Matters* 1, 4 (2015), 4–12.
- Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., and Taylor, J. Industry-scale knowledge graphs: lessons and challenges. *Commun. ACM* 62, 8 (Aug. 2019), 36–43.
- Patel-Schneider, P. and Horrocks, I. Position paper: A comparison of two modelling paradigms in the Semantic Web. In *Proceedings of the 15th ACM Intern. Conf. World Wide Web*, (Edinburgh, Scotland, May 23–26, 2006). L. Carr, D. De Roure, A. Iyengar, C.A. Goble, and M. Dahlin (Eds.), 3–12.
- Rietveld, L., Beek, W., and Schlobach, S. LOD lab: Experiments at LOD scale. In *Proceedings of the 14th Intern. Semantic Web Conf. (Bethlehem, PA, USA, Oct. 11–15, 2015)*. M. Arenas et al. (Eds.). LNCS 9367. Springer, 339–355.
- Sabou, M., Biffl, S., Einfalt, A., Krammer, L., Kastner, W., and Ekaputra, F. Semantics for cyber-physical systems: A cross-domain perspective. *Semantic Web* 11, 1 (2020), 115–124.
- Sandhaus, E. Abstract: Semantic technology at the New York Times: Lessons learned and future directions. In *Proceedings of the 9th Intern. Semantic Web Conf. (Shanghai, China, Nov. 7–11, 2010)*. P.F. Patel-Schneider et al. (Eds.), LNCS 6497. Springer, 355.
- Schreiber, G. and Raimond, Y. (Eds.). *RDF 1.1 Primer*. W3C Working Group Note (June 24, 2014); <http://www.w3.org/TR/rdf11-primer/>.
- Schulz, S., Sunitisivaraporn, B., Baader, F., and Boeker, M. SNOMED reaching its adolescence: Ontologists' and logicians' health check. I. *J. Medical Informatics* 78, Supp. 1 (2009), S86–S94.
- Sporny, M., Longley, D., Kellogg, G., Lanthaler, M., and Lindström, N. JSON-LD 1.0. A JSON-based Serialization for Linked Data. W3C Recommendation (Jan. 16, 2014); <http://www.w3.org/TR/jsonld/>.
- The Gene Ontology Consortium. The Gene Ontology Project in 2008. *Nucleic Acids Research* 36 (Database issue) (2008), D440–D444.
- The W3C SPARQL Working Group (Ed.). *SPARQL 1.1 Overview*. W3C Recommendation (Mar. 21, 2013); <http://www.w3.org/TR/sparql11-overview/>.
- Thieblin, E., Haemmerle, O., Hernandez, N., and Santos, C. Survey on complex ontology matching. *Semantic Web* (2020), 689–727.
- Tudorache, T., Nyulas, C., Noy, N., and Musen, M. Using Semantic Web in ICD-11: Three years down the road. In *Proceedings of the 12th Intern. Semantic Web Conf. (Sydney, NSW, Australia, Oct. 21–25, 2013)*. H. Alani et al. (Eds.). LNCS 8219. Springer, 195–211.
- Vrandečić, D. and Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57, 10 (Oct. 2014), 78–85.
- Weaver, J. and Tarjan, P. Facebook linked data via the graph API. *Semantic Web* 4, 3 (2013), 245–250.

**Pascal Hitzler** is a professor and endowed Lloyd T. Smith Creativity in Engineering Chair and director of the Center for Artificial Intelligence and Data Science in the Department of Computer Science at Kansas State University, Manhattan, KS, USA.

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



Watch the author discuss this work in the exclusive *Communications* video. <https://caom.acm.org/videos/semantic-web>

## Synthesizing the emerging directions of research at the intersection of differential privacy and cryptography.

BY SAMEER WAGH, XI HE, ASHWIN MACHANAVAJHALA, AND PRATEEK MITTAL

# DP-Cryptography: Marrying Differential Privacy and Cryptography in Emerging Applications

ON FEB 15, 2019, John Abowd, chief scientist at the U.S. Census Bureau, announced the results of a *reconstruction attack* that they proactively launched using data released under the 2010 Decennial Census.<sup>19</sup> The decennial census released billions of statistics about individuals like “how many people of the age 10–20 live in New York City” or “how many people live in four-person households.” Using only the data publicly released in 2010, an internal team was able to correctly reconstruct records of address (by census block), age, gender, race, and ethnicity for 142 million

people (about 46% of the U.S. population), and correctly match these data to commercial datasets circa 2010 to associate personal-identifying information such as names for 52 million people (17% of the population).

This is not specific to the U.S. Census Bureau—such attacks can occur in any setting where statistical information in the form of deidentified data, statistics, or even machine learning models are released. That such attacks are possible was predicted over 15 years ago by a seminal paper by Irit Dinur and Kobbi Nissim<sup>12</sup>—releasing a sufficiently large number of aggregate statistics with sufficiently high accuracy provides sufficient information to reconstruct the underlying database with high accuracy. The practicality of such a large-scale reconstruction by the U.S. Census Bureau underscores the grand challenge that public organizations, industry, and scientific research faces: How can we safely disseminate results of data analysis on sensitive databases?

An emerging answer is *differential privacy*. An algorithm satisfies differential privacy (DP) if its output is insensitive to adding, removing or changing one record in its input database. DP is considered the “gold standard” for privacy for a number of reasons. It provides a persuasive mathematical proof of privacy to individuals with several rigorous interpretations.<sup>25,26</sup> The DP guarantee is composable and repeating

### » key insights

- **Local Differential Privacy is increasingly being embraced as the primary model of deployment of differential privacy, albeit at a heavy accuracy cost.**
- **Cryptographic primitives can help bridge the utility gap between systems deployed in the local differential privacy model and standard differential privacy model, but the increased utility may come at the cost of performance.**
- **DP-cryptographic primitives, which are relaxed notions of cryptographic primitives that leak differentially private outputs, permit implementations that are orders of magnitude faster than the regular primitives.**





invocations of differentially private algorithms lead to a graceful degradation of privacy. The U.S. Census Bureau was the first big organization to adopt DP in 2008 for a product called OnTheMap,<sup>29</sup> and subsequently there have been deployments by Google, Apple, Microsoft, Facebook, and Uber.<sup>2,11,17,18,36</sup>

DP is typically implemented by collecting data from individuals in the clear at a trusted data collector, then applying one or more differentially private algorithms, and finally releasing the outputs. This approach, which we call *standard differential privacy (SDP)*, works in cases like the U.S. Census Bureau where there is a natural trusted data curator. However, when Google wanted to monitor and analyze the Chrome browser properties of its user base to detect security vulnerabilities, they chose a different model called *local differential privacy (LDP)*. In LDP, individuals perturb their records *before* sending them to the server, obviating the need for a trusted data curator. Since the server only sees perturbed records, there is no centralized database of sensitive information that is susceptible to an attack or subpoena requests from governments. The data that Google was collecting—browser fingerprints—uniquely identify individuals. By using LDP, Google was not liable to storing these highly identifying user properties. Due to these attractive security properties, a number of real-world applications of DP in the industry—Google’s RAPPOR,<sup>17</sup> Apple Diagnostics<sup>2</sup> and Microsoft Telemetry<sup>11</sup>—embrace the LDP model.

However, the improved security properties of LDP come at a cost in terms of utility. DP algorithms hide the presence or absence of an individual by adding noise. Under the SDP model, counts over the sensitive data, for example, “number of individuals who use the bing.com search engine,” can be released by adding a noise independent of the data size. In the LDP model, noise is added to *each individual record*. Thus, answering the same count query requires adding  $O(\sqrt{N})$  error (Theorem 2.1 from Chen et al.<sup>10</sup>) for the same level of privacy, where  $N$  is the number of individuals participating in the statistic. In other words, under the LDP model, for a database of a billion people, one can only learn properties that



**When used in practice, practical trust assumptions are made that enable the deployment of differential privacy-based systems.**



are common to at least 30,000 people  $O(\sqrt{N})$ . In contrast, under SDP, one can learn properties shared by as few as a 100 people ( $O(1)$  including constants<sup>15</sup>). Thus, the LDP model operates under more practical trust assumptions than SDP, but as a result incurs a significant loss in data utility. In this work, we review literature in this domain under two categories:

► **Cryptography for DP:** We review a growing line of research that aims to use cryptographic primitives to bridge the gap between SDP and LDP. In these solutions, the trusted data curator in SDP is replaced by cryptographic primitives that result in more practical trust assumptions than the SDP model, and better utility than under the LDP model. Cryptographic primitives such as anonymous communication and secure computation have shown significant promise in improving the utility DP implementations while continuing to operate under the practical trust assumptions that are accepted by the security community.

► **DP for cryptography:** Differential privacy is typically applied to settings that involve complex analytics over large datasets. Introducing cryptographic primitives results in concerns about the feasibility of practical implementations at that scale. This has given rise to a second line of work that employs differential privacy as a tool to speed up cryptographic primitives, thereby pushing the frontiers of their practical deployments. While the original cryptographic primitives are defined with respect to perfect privacy, under differential privacy, it is OK to learn distributional information about the underlying dataset. We explore in depth the following cryptographic primitives: secure computation and secure communication and show how in the context of differential privacy one can build “leaky” but efficient implementations of these primitives.

These lines of work both reflect exciting directions for the computer science community. We begin by giving a brief technical introduction to DP. We then discuss the “Cryptography for DP” and “DP for cryptography” paradigms. Finally, we provide concrete ideas for future work as well as open problems in the field through the lens of combining differential privacy and cryptography.

## Differential Privacy

Differential privacy<sup>13</sup> is a state-of-the-art privacy metric for answering queries from statistical databases while protecting individual privacy. Since its inception, there has been considerable research in both the theoretical foundations<sup>12,14</sup> as well as some real-world DP deployments.<sup>2,17</sup> The rigorous mathematical foundation and the useful properties of DP have led to an emerging consensus about its use among the security and privacy community.

**DP definition.** Informally, the privacy guarantees of differential privacy can be understood as follows: Given any two databases, otherwise identical except one of them contains random data in place of data corresponding to any *single* user, differential privacy requires that the response mechanism will behave approximately the same on the two databases. Formally,

**Definition 1.** Let  $M$  be a randomized mechanism that takes a database instance  $D$  and has a range  $O$ . We say  $M$  is  $(\epsilon, \delta)$ -differentially private, if for any neighboring databases  $(D_1, D_2)$  that differ in the data of a single user, and for any  $S \subseteq O$ , we have

$$\Pr[M(D_1) \in S] \leq e^\epsilon \Pr[M(D_2) \in S] + \delta$$

DP enjoys some important properties that make it a useful privacy metric. First, the privacy guarantees of DP have been thoroughly studied using various metrics from statistics and information theory such as hypothesis testing and Bayesian inference.<sup>25,26</sup> Thus, the semantic meaning of its privacy guarantees is well understood. DP also has a number of composition properties which enable the analysis of privacy leakage for complex algorithms. In particular, sequential composition addresses the impossibility result by Dinur and Nissim<sup>12</sup> and quantifies the degradation of privacy as the number of sequential accesses to the data increases. The post-processing theorem (a special case of sequential composition) ensures the adversary cannot weaken the privacy guarantees of a mechanism by transforming the received response. The end-to-end privacy guarantee of an algorithm over the entire database can thus be established using the above composition theorems and more advanced theorems.<sup>15</sup>

## Differentially private mechanisms.

Next, we review two classic differentially private mechanisms—the Laplace mechanism and the Randomized Response mechanism—with the following scenario: A data analyst would like to find out how many users use drugs illegally. Such a question would not elicit any truthful answers from users and hence we require a mechanism that guarantees (a) response privacy for the users and (b) good utility extraction for the data analyst.

**Laplace mechanism:** The Laplace mechanism<sup>13</sup> considers a trusted data curator (SDP model) who owns a table of  $N$  truthful records of users, for example, each record indicates whether a user uses drugs illegally. If a data analyst would like to learn how many users use drugs illegally, the data curator (trusted) computes the true answer of this query and then perturbs it with a random (Laplace distributed) noise that is sufficient to provide privacy. The magnitude of this noise depends on the largest possible change on the query output—also known as the sensitivity of the query—if the data corresponding to a single user is changed.

**Randomized response mechanism:** Randomized response was first introduced by Warner in 1965 as a research technique for survey interviews. It enabled respondents to answer sensitive questions (about topics such as

sexuality, drug consumption) while maintaining the confidentiality of their responses. An analyst interested in learning aggregate information about sensitive user behavior would like to query this function on a database that is *distributed* across  $N$  clients with each client having its own private response  $x_1, \dots, x_N$ . Instead of releasing  $x_i$  directly, the clients release a perturbed version of their response  $y_i$ , thus maintaining response privacy. The analyst collects these perturbed responses and recovers meaningful statistics using reconstruction techniques.

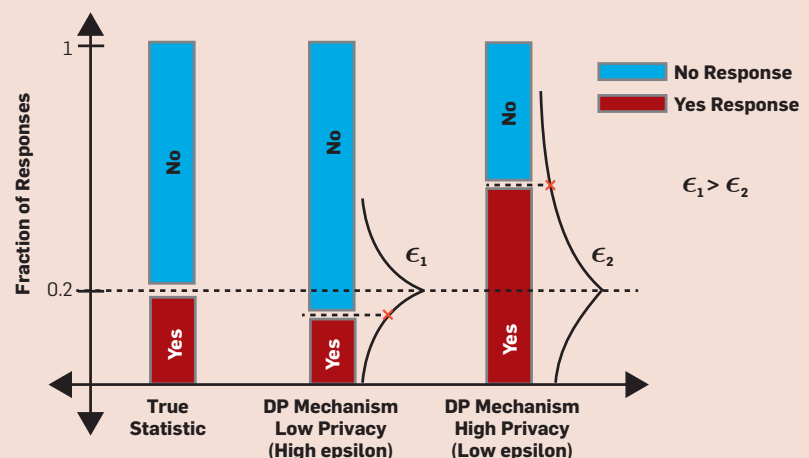
Both these approaches have gained popularity in many applications of differential privacy due to their simplicity as well as the rigorous privacy guarantee on user data. Figure 1 shows the behavior of DP mechanisms for two different privacy values in reference to the true statistic. A less private response results in a more accurate query result while a more private response results in a less accurate query result.

## Cryptography for Differential Privacy

By itself, DP is a guarantee on a mechanism and hence is “independent” of the deployment scenario. However, when used in practice, practical trust assumptions are made that enable the deployment of differential privacy-based systems. Here,

**Figure 1. Differentially private mechanisms randomize query response to achieve privacy.**

If the true response to a query such as “What fraction of users use drugs illegally?” was 20%, then a high privacy response mechanism (low  $\epsilon$  value) will add a lot of noise yielding low utility. On the contrary, if a low privacy response mechanism was used (high  $\epsilon$  value), the response will be very close to 20% yielding high utility.



we consider two popular deployment scenarios for differential privacy—Standard Differential Privacy (SDP, graphically represented in Figure 2D) and Local Differential Privacy (LDP, graphically represented in Figure 2A). SDP relies on the need for a trusted data aggregator who follows the protocol. However, in practice, a trusted data aggregator may not always exist. LDP, on the other hand, does not require a trusted data aggregator.<sup>a</sup> With the advent of privacy regulations, such as GDPR and FERPA, large organizations such as Google increasingly embrace the LDP model thereby avoiding the liability of storing such sensitive user data. This approach also insures data collectors from potential theft or subpoenas from the government. For these reasons, LDP is frequently a more attractive deployment scenario. However, the utility of the statistics released in LDP is poorer than that in SDP. Consequently, there is a gap in the trust assumptions and the utility achieved by mechanisms in SDP and LDP: high trust assumptions, high utility in SDP and lower trust assumptions, lower utility in LDP. We ask the following question:

a Differentially private federated learning is simply a special case of the LDP deployment scenario.

Can cryptographic primitives help in bridging the gap that exists between mechanisms in the SDP model and the LDP model?

An emerging direction of research has been to explore the use of cryptography to bridge the trust-accuracy gap and obtain the best of both worlds: high accuracy without assuming trusted data aggregator. We explore in depth two concrete examples of the role of cryptography in bridging this gap—anonymous communication, and secure computation and encryption.

**Key challenges.** There exists a big gap in the accuracy and trust achieved by known mechanisms in the SDP setting with a trusted data curator (Figure 2D) and LDP without such a trusted curator (Figure 2A). Achieving the utility as in the SDP setting while operating under practical trust assumptions such as those in LDP has proven to be a tough challenge. Cryptographic primitives show promise in solving this challenge.

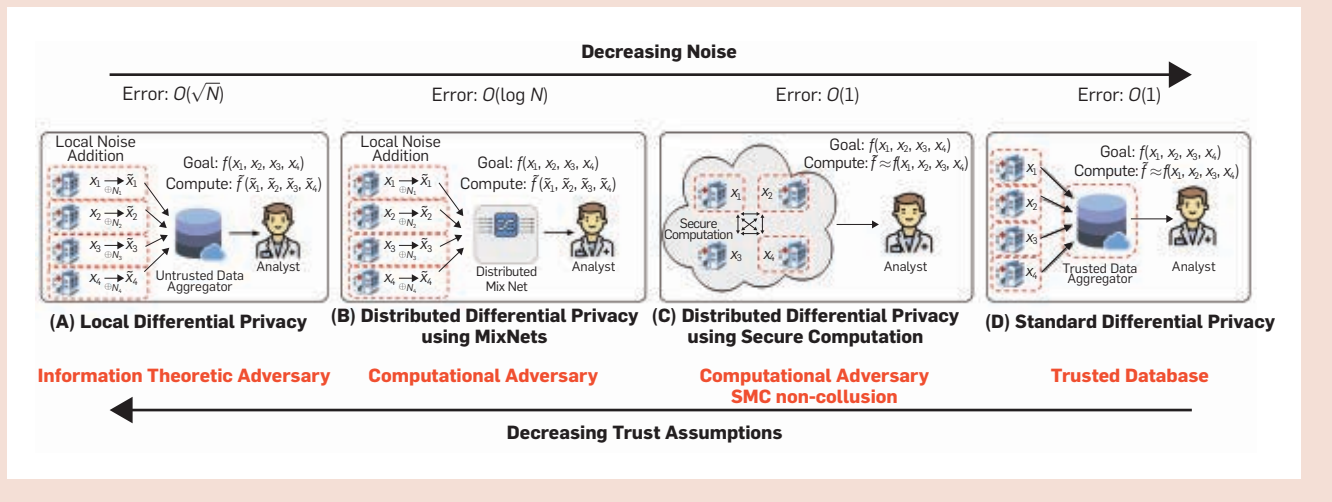
**Improve accuracy via anonymous communication.** In LDP, each data owner independently perturbs their own input (for example, using the randomized response mechanism) before the aggregation on an untrusted server. This results in a large noise in the final output,  $O(\sqrt{N})$  for the case of statistical

counting queries.<sup>10</sup> Applications such as Google’s RAPPOR,<sup>17</sup> Apple Diagnostics,<sup>2</sup> and Microsoft Telemetry,<sup>11</sup> which use this LDP deployment model operate under more practical trust assumptions yet suffer from poor accuracy/utility. Recent works<sup>8,10,16</sup> show the use of an anonymous communication channel can help improve the accuracy of statistical counting query for LDP and thereby eliminate the need for a trusted data curator. We will use one of these systems called Prochlo<sup>8,16</sup> to illustrate the key idea of how anonymous communication can help improve the accuracy of such applications.

**Case Study: Prochlo.** Anonymous communication channels, first proposed by Chaum in 1981,<sup>9</sup> are systems that enable a user to remain unidentifiable from a set of other users (called the anonymity set). A larger anonymity set corresponds to a greater privacy guarantee. Examples of such systems include Mixnets, which use proxies to mix communications from various users. In order to circumvent the limitations of LDP, Google explored the use of an anonymous communication channel to improve the accuracy of queries under DP. The proposed technique is called Prochlo<sup>8,16</sup> and it consists of three steps as shown in Figure 2B: Encode, Shuffle, and Analyze (ESA).

**Figure 2. Various deployment scenarios of differential privacy and the underlying trust assumptions in each of them.**

(D) Standard Differential Privacy (SDP) assumes a trusted database, and is thus able to achieve high accuracy, such as,  $O(1)$  error. (A) Local Differential Privacy (LDP) on the other hand, does not rely on the use of a trusted database but achieves lower accuracy, that is,  $O(\sqrt{N})$  error. The goal is to achieve utility of the SDP setting while operating under more practical assumptions such as the LDP setting (that is, no trusted database). (B) and (C) show how different cryptographic primitives can be used to improve the utility of DP deployments under such practical assumptions.







The first encoding step is similar to LDP where data owners randomize their input data independently. The second step uses an anonymous communication channel to collect encoded data into batches during a lengthy time interval and shuffles this data to remove the linkability between the output of the communication channel and the data owners. Last, the anonymous, shuffled data is analyzed by a data analyst.

The shuffling step is the crucial link in achieving anonymous communication by breaking linkability between the user and their data. This step strips user-specific metadata such as time stamps or source IP addresses, and batches a large number of reports before forwarding them to data analysts. Additional thresholding in this step will discard highly unique reports (for example, a long API bit-vector) to prevent attackers with sufficient background information from linking a report with its data owner. Hence, attacks based on traffic analysis and longitudinal analysis can be prevented, even if a user contributes to multiple reports. Prochlo implements this shuffling step using trusted hardware as proxy servers to avoid reliance on external anonymity channel. Furthermore, this shuffling step can amplify the privacy guarantee of LDP and hence improves the accuracy of the analysis, even when there is a single invocation from a user. We will next show the intuition for this use case.

*Accuracy improvement.* To illustrate how anonymous communication can help improve accuracy, let us look at a simple example of computing the sum of boolean values from  $N$  data owners,  $f: \Sigma_{i=1}^N x_i$ , where  $x_i \in \{0, 1\}$ . In LDP, each data owner reports a random bit with probability  $p$  or reports the true bit with probability  $1 - p$  to achieve  $\epsilon$ -LDP. When using additional anonymous communication channels, the data owners can enhance their privacy by hiding in a large set of  $N$  values, since the attackers (aggregator and analyst) see only the anonymized set of reports  $\{\tilde{x}_1, \dots, \tilde{x}_N\}$ . The improved privacy guarantee can be shown equivalent to a simulated algorithm that first samples a value  $s$  from a binomial distribution  $B(N, p)$  to simulate the number of data owners who report a random bit, and then samples a subset



**Cryptographic primitives provide strong privacy guarantees. However, deployment of certain cryptographic primitives in practical systems is limited due to the large overhead of these primitives.**



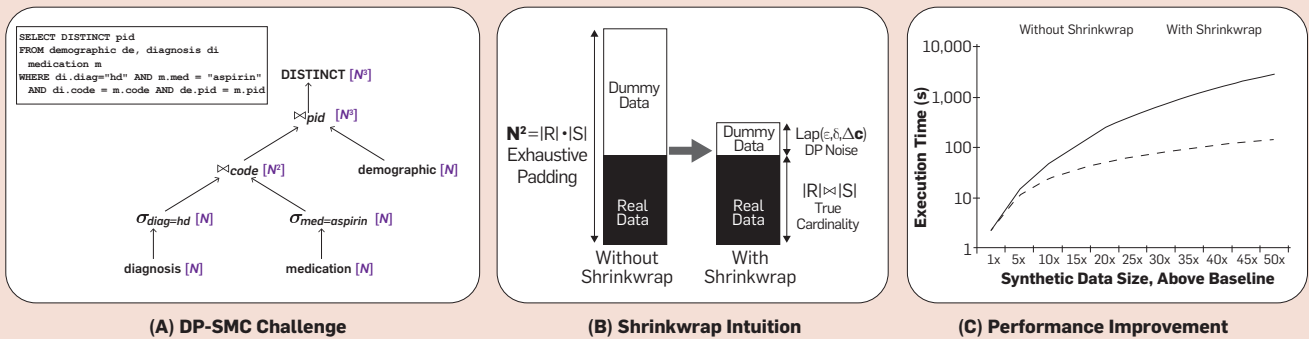
of responses for these  $s$  data owners from  $\{\tilde{x}_1, \dots, \tilde{x}_N\}$ . The randomness of these sampling processes can amplify the privacy parameter based on a well-studied sub-sampling argument.<sup>3,23</sup> Therefore, given the value of the privacy parameter, the required noise parameter can be scaled down and hence the corresponding error can be reduced to  $O(\sqrt{\log N})$ . However, these bounds depend on the specific deployment scenarios. For instance, it is shown in Balle et al.<sup>4</sup> that anonymous communication with a single message per data owner cannot yield expected error less than  $O(N^{1/6})$ . On the other hand, works such as Balle<sup>5</sup> and Kasiviswanathan et al.<sup>26</sup> show that with a constant number of messages per data owner, it is possible to reduce the error for real-valued DP summation to  $O(1)$ . Note that these accuracy improvements assume that there is no collusion between the analyst and the anonymous communication, otherwise, the privacy guarantee will fall back to the same as LDP.

In reference to Figure 2, these works demonstrate the improvement in going from Figure 2A to Figure 2B showing a trade-off between accuracy and trust assumptions.

**Improve trust via encryption and secure computation.** SDP requires the use of a trusted data aggregator to achieve high accuracy. A number of works have explored the use of encryption and secure computation to eliminate the need for this trusted data aggregator.<sup>1,6,33</sup> The key challenge here is to maintain the same level of accuracy as in SDP. We will use one of these proposed systems called DJoin to demonstrate the use of secure computation to enable high accuracy computation without the need for a trusted data aggregator. There is a complementary synergy between secure computation and DP and thus their combination achieves a strong privacy protection. For instance, secure computation ensures all parties learn only the output of the computation but nothing else while DP bounds the information leakage of individuals in the output of the computation, resulting in a system that is better than the use of secure computation or DP alone.

*Case Study: DJoin.* Consider a simple setting where two parties would like to compute the intersection size of their

**Figure 3. (A) Exhaustive padding of intermediate results in an oblivious query evaluation; (B) Effect of Shrinkwrap on intermediate result sizes when joining tables R and S; (C) Aspirin count with synthetic data scaling. Executed using Circuit model.  $\epsilon = 0.5, \epsilon = .00005$ .**



data while preserving DP for both datasets. If each party does not trust each other, how can we ensure a constant additive error as if they trust each other? It is well known that the lower bound for the error of this query is  $\sqrt{N}$ , where  $N$  is the data size of each party,<sup>30</sup> if we want to ensure the view of each party satisfies differential privacy. However, if we assume both parties are computationally bounded, a constant additive error can be achieved.

DJoin<sup>33</sup> offers a concrete protocol for achieving DP under this assumption. This protocol applies private set-intersection cardinality technique to privately compute the noisy intersection set of the two datasets. First, party A defines a polynomial over a finite field whose roots are the elements owned by A. Party A then sends the homomorphic encryptions of the coefficients to party B, along with its public key. Then the encrypted polynomial is evaluated at each of Party B's inputs, followed by a multiplication with a fresh random number. The number of zeros in the results is the true intersection size between A and B. To provide DP, party B adds a number of zeros (differentially private noise of  $O(1)$  independent of data size) to the results and sends the randomly permuted results back to party A. Party A decrypts the results and counts the number of zeros. Party A also adds another copy of differentially private noise to the count and sends the result it back to party B. In other words, both parties add noise to their inputs to achieve privacy. However, the final protocol output has only an error of  $O(1)$ , which is the same as the SDP setting.

*Trust improvement.* Using secure computation and encryptions achieves

a constant additive error like SDP and prevents any party from seeing the other party's input in the clear. However, this requires an additional assumption of all parties being computationally bounded in the protocol. Hence, the type of DP guarantee achieved in DJoin is known as computational differential privacy.<sup>32</sup> In addition, most of the existing protocols consider honest-but-curious adversaries who follow the protocol specification or consider malicious adversaries with an additional overhead to enforce honest behavior, that is, verify that the computation was performed correctly.

In reference to Figure 2, these works demonstrate the improvement in going from Figure 2D to Figure 2C eliminating the need for a trusted data aggregator.

**Differential Privacy for Cryptography**

As we discussed earlier, cryptographic primitives show promise in bridging the utility gap between SDP and LDP. However, the large overhead of implementing these conventional cryptographic primitives forms a bottleneck for the deployment of such systems. This motivates the need to enhance the performance of such cryptographic primitives. We ask the following question:

**“Can we formulate leaky versions of cryptographic primitives for enhancing system performance while rigorously quantifying the privacy loss using DP?”**

DP-cryptographic primitives<sup>7,37,38</sup> are significant for two reasons. First, since the final privacy guarantees of such systems are differential privacy, it is natural to relax the building blocks

such as cryptographic primitives to provide differentially private guarantees. Secondly, the composability properties of DP allow for rigorous quantification of the privacy of the end-to-end system. We showcase benefits of “DP-cryptographic” systems through two detailed case studies on secure computation and secure communication.

*Key challenges.* Cryptographic primitives provide strong privacy guarantees. However, deployment of certain cryptographic primitives in practical systems is limited due to the large overhead of these primitives. Relaxing the privacy guarantees in a manner that is amenable to rigorous quantification is difficult and differential privacy can be well utilized to provide a solution to this problem to improve performance overhead.

**Improve performance of cryptographic computation primitives.**

Cryptographic computation primitives such as Fully Homomorphic Encryption (FHE) and secure Multi-Party Computation (MPC) enable private computation over data. Over the past few years, there has been tremendous progress in making these primitives practical—a promising direction is MPC, which allows a group of data owners to jointly compute a function while keeping their inputs secret. Here, we show the performance improvement on MPC based private computation, in particular, *differentially private query processing*.

*Case Study: Shrinkwrap.* Shrinkwrap<sup>7</sup> is a system that applies DP throughout an SQL query execution to improve performance. In secure computation, the computation overheads depend on the largest possible data

size so that no additional information is leaked. For example, two parties would like to securely compute the answer for the SQL query shown in Figure 3A. This query asks for the number of patients with heart disease who have taken a dosage of “aspirin.” Figure 3A expresses this query as a directed acyclic graph of database operators. For example, the first filter operator takes  $N$  records from the two parties and outputs an intermediate result that has patients with heart disease (hd). To hide the selectivity (fraction of records selected) of this operator, the baseline system must pad the intermediate result to its maximum possible size, which is the same as the input size. Exhaustive padding will also be applied to the intermediate output of the two joins and result in an intermediate result cardinality of  $N^3$  and a high-performance overhead. However, if the selectivity of the filter is  $10^{-3}$ , cryptographic padding adds a  $1000\times$  overhead. Is there a way to pad fewer dummies to the intermediate result while ensuring a provable privacy guarantee?

Shrinkwrap helps reduce this overhead by padding each intermediate output of the query plan to a differentially private cardinality rather than to the worst case. As shown in Figure 3B, without Shrinkwrap, the output of a join operator with two inputs, each of size  $N$  is padded to a size of  $N^2$ . With Shrinkwrap, the output is first padded to the worst size and the output is

sorted such that all the dummies are at the end of the storage. This entire process is executed obviously. Then Shrinkwrap draws a non-negative integer value with a general Laplace mechanism<sup>7</sup> and truncates the storage at the end. This approach reduces the input size of the subsequent operators and thereby their I/O cost. We can see from Figure 3C that Shrinkwrap provides a significant improvement in performance over the baseline without DP padding for increasing database sizes.

The relaxed privacy in the secure computation of Shrinkwrap can be quantified rigorously<sup>7</sup> using computational differential privacy. Assuming all parties are computationally bounded and work in the semi-honest setting, it can be shown that data owners have a computational differentially private view over the input of other data owners; when noisy answers are returned to the data analyst, the data analyst has a computational differentially private view over the input data of all the data owners.

**Improve performance of cryptographic communication primitives.**

Anonymous communication systems aim to protect user identity from the communication recipient and third parties. Despite considerable research efforts in this domain, practical anonymous communication over current Internet architecture is proving to be a challenge. Even if the message contents are encrypted, the packet metadata is

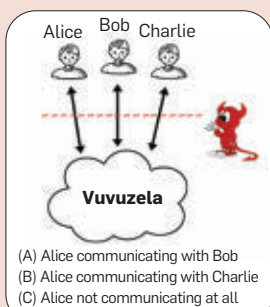
difficult to hide. On one end, systems such as Dissent<sup>39</sup> offer strong privacy guarantees yet can scale only to a limited number of participants. On the other end, practical deployed systems such as Tor are vulnerable to traffic analysis and other attacks, limiting their use due to the non-rigorous nature of their privacy guarantees. We will show a case study that uses DP to reduce the communication cost while offering rigorous privacy guarantee. We denote this primitive differentially private anonymous communication.

*Case Study: Vuvuzela.* Vuvuzela<sup>37</sup> is an anonymous communication system that uses DP to enable a highly scalable system with relaxed yet rigorously quantified privacy guarantees. Vuvuzela provides indistinguishable traffic patterns to clients who are actively communicating with other clients, and clients who are not communicating with anyone. In reference to Figure 4, an adversary is unable to distinguish the following three scenarios: Alice not communicating; Alice communicating with Bob; and, Alice communicating with Charlie. In each of the scenarios, a Vuvuzela client’s network traffic appears indistinguishable from the other scenarios.

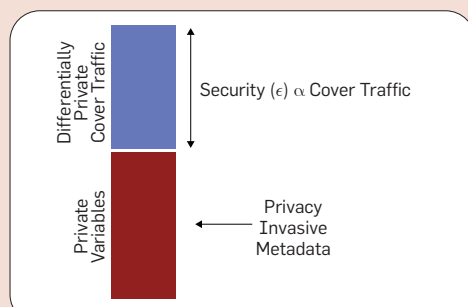
Vuvuzela employs a number of servers  $S_1, \dots, S_n$  where at least one of the servers is assumed to be honest. Clients send (and receive) messages to (and from) the first server, which in turn is connected to the second server and so

**Figure 4. Vuvuzela is a secure messaging system.**

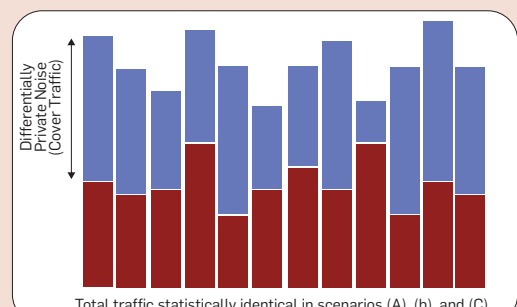
An adversary who can observe and tamper with all network traffic cannot distinguish whether Alice is messaging Bob, Charlie, or is simply not communicating. Vuvuzela uses differential privacy to add noise and mask the privacy invasive metadata, thereby provably hiding information about user communication patterns. Vuvuzela achieves a throughput of 68,000 messages per second for a million users scaling linearly with number of users.



**(A) Schematic for Vuvuzela**



**(B) Performance improvement of Vuvuzela**



**(C) Security proportional to cover traffic**



on. The client creates a layered encryption of its message  $m$ , that is,  $\text{Enc}_{s_1}(\dots \text{Enc}_{s_n}(m))$ , where  $\text{Enc}_S(\cdot)$  is the encryption under the key of server  $S$ . The clients leave messages at virtual locations in a large space of final destinations (called dead drops), where the other legitimate client can receive it. To hide if a client is communicating or not, a client not in an active conversation makes fake requests to appear indistinguishable from a client in an active conversation. If two clients are in active conversation, they exchange messages via the same random dead drop.

Vuvuzela's threat model assumes at least one server is honest and the adversary is a powerful network level adversary (observing all network traffic) potentially corrupting all other servers.<sup>b</sup> The only computation hidden from the adversary is the local computation performed by the honest server which unlinks users' identifiers from the dead drops and adds cover (dummy) traffic. As a consequence, the adversary can only observe the number of single or double exchange requests at the dead drop locations. Each Vuvuzela server adds cover traffic using a Laplace distribution to randomize the number of single dead drops and the number of double dead drops, which is observable by the adversary. Such random cover traffic addition along with the assumption of at least one honest server provides DP guarantees for the observed variables. In other words, Vuvuzela adds noise (cover network traffic) to the two observables (by the adversary) viz. the number of dead drops with one exchange request, and the number of dead drops with two exchange requests, thereby providing communication privacy to clients. This privacy relaxation enables Vuvuzela to scale to a large number of users—it can achieve a throughput of 68,000 messages per second for a million users. Systems such as Stadium,<sup>35</sup> and Karaoke<sup>27</sup> further improve upon Vuvuzela and scale to even larger sets of users.

**Limitations of differentially private cryptography.** We caution readers against careless combinations of differential privacy and cryptographic primitives. First, the limitations of both DP as well as cryptographic primitives apply to DP

cryptographic primitives. For instance, an open question is deciding an appropriate level for the privacy budget. Most applications that utilize DP to improve the performance of cryptographic systems involve a trade-off between the level of privacy achieved and the performance of the systems. More generally, differentially private cryptographic systems open up new trade-offs in a privacy-performance-utility space. For instance, in the case of Shrinkwrap, weaker privacy guarantee directly leads to lower performance overhead (privacy performance trade-off while keeping the accuracy level of the query answer constant). On the other hand, systems such as RAPPOR allow for approximate computation of statistics and primarily provide a privacy-utility trade-off. Second, designers need to carefully consider the suitability of these hybrid techniques in their applications as these combinations involve more complex trust assumptions and hence a more complicated security analysis. We remind the reader that while proposing newer DP systems for cryptography, it is imperative to understand the meaning of the privacy guarantees for the application in context. In other words, differentially privacy for cryptography may not be the right thing to do in all cases; however, it is well motivated when the goal is to build a differentially private system. Finally, composition results, which bound the privacy loss for a sequence of operations need to be independently studied.

### Discussion and Open Questions

Here, we provide directions for future work highlighting important and emerging open questions in the field. We discuss open challenges in deploying differential privacy in the real world—realistic datasets, alternative models and trust assumptions, and other DP-cryptographic primitives. Finally, we caution readers against callous combinations of differential privacy and cryptography.

**Differential privacy frameworks—SDP, LDP, and beyond.** Over the past decade, there has been significant progress in enabling applications in the standard differential privacy model. For instance, there have been research efforts in attuning DP to handle realistic challenges such as multi-dimensional and complex data—involving graphs,

time series, correlated data.<sup>24,28</sup> Similarly, there has been work in designing a tailored DP mechanism that is optimized for particular application setting to achieve good accuracy.<sup>22,31</sup> Prior work has explored combinations of sequential and parallel composition, dimensionality reduction, and sensitivity bound approximations to achieve good accuracy in the SDP model. However, much work needs to be done in adapting state-of-the-art techniques in SDP to more complex deployment scenarios such as LDP. For instance, an open question is the following:

**“Is there an algorithm that can efficiently search the space of DP algorithms in the LDP setting for the one that answers the input query with the best accuracy?”**

Research advances have demonstrated such mechanisms for the SDP model,<sup>22,31</sup> however, the discovery of such mechanisms in the LDP setting remains an open question. On a similar note, it is unclear how nuanced variants of DP that have been proposed to handle these more complex databases<sup>24,28</sup> in the SDP setting translate into LDP or more complex deployment settings.

### Differential privacy in practice—Trust assumptions vs accuracy gap.

We have seen how deployments of DP that differ in the trust assumptions provide approximately the same privacy guarantee, but with varying levels of accuracy. In particular, we looked at two popular deployment scenarios: SDP and LDP. There exist other trust assumptions that we have not covered in this article in detail. For instance, Google's recently proposed Prochlo system<sup>8</sup> uses trusted hardware assumptions to optimize utility of data analytics. On a similar note, Groce et. al.<sup>21</sup> consider yet another model—where the users participating are malicious. This is the first work to explore a malicious adversarial model in the context of DP and the development of better accuracy mechanisms for such a model is an open research question. More concretely, we can ask:

**“What other models of deployment of differential privacy exist and how**

<sup>b</sup> Even Tor, a practical anonymous communication system, does not protect against such network level adversaries.<sup>34</sup>

## do we design mechanisms for them? Can other technologies such as MPC, FHE, trusted hardware opens up new opportunities in mechanism design?”

An interesting theoretical question is to characterize the separation between different trust models in terms of the best accuracy achievable by a DP algorithm under that model. For instance, McGregor et. al.<sup>30</sup> provide separation theorems, that is, gaps in achievable accuracy between (information-theoretic) differential privacy and computational differential privacy for two-party protocols. We ask:

In the Mixnets model (Figure 2B), what is the lower bound on the error for aggregate queries over relational transformations (like joins and group-by) over the data records? An example of such an aggregate is the degree distribution of a graph that reports the number of nodes with a certain degree.

**Relaxing cryptographic security via DP:** The emerging paradigm of leaky yet differentially private cryptography leads to a number of open questions for the research community. So far, the research community has explored the intersection of differential privacy and cryptographic primitives in limited contexts such as ORAM, MPC, and anonymous communication. However, there exists a broader opportunity to explore the trade-offs of DP cryptographic primitives in contexts such as program obfuscation, zero-knowledge proofs, encrypted databases, and even traffic/protocol morphing. Here, we can ask:

## “What other cryptographic primitives can benefit in performance from a privacy relaxation quantified rigorously using differential privacy? How can we design such relaxed primitives?”

In the context of differentially private data analysis, there is a trade-off between privacy and utility. In the context of differentially private cryptographic primitives and resulting applications, there is a broader trade-off space between privacy, utility, and performance. Another open question is:

## “What lower bounds exist for overhead of cryptographic

## primitives when the privacy guarantees are relaxed using DP?”

Another challenge is how to design optimized protocols that achieve desired trade-offs in the new design space of differentially private cryptography. The trade-off space between privacy, utility, and performance is non-trivial, especially for complex systems. An interesting research question is:

## “How to correctly model the trade-off space of real systems so that system designers can decide whether it is worth sacrificing some privacy or utility for a better performance?”

### References

1. Agarwal, A., Herlihy, M., Kamara, S., and Moataz, T. Encrypted Databases for Differential Privacy. *IACR Cryptology ePrint Archive*, 2018.
2. Apple is using Differential Privacy to help discover the usage patterns of a large number of users without compromising individual privacy; <https://apple.co/3ctHYkw>
3. Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *Advances in Neural Information Processing Systems 31*. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Curran Associates, Inc., 2018, 6277–6287.
4. Balle, B., Bell, J., Gascon, A., and Nissim, K. The privacy blanket of the shuffle model. In *Proceedings of the Annual Intern. Cryptology Conf.* Springer, 2019, 638–667.
5. Balle, B., Bell, J., Gascon, A., and Nissim, K. Private summation in the multi-message shuffle model. *arXiv preprint arXiv:2002.00817*, 2020.
6. Bater, J., Elliott, G., Eggen, C., Goel, S., Kho, A., and Rogers, J. SMCQL: Secure querying for federated databases. In *Proceedings of the VLDB Endowment 10*, 6 (2017), 673–684.
7. Bater, J., He, X., Ehrlich, W., Machanavajhala, A., and Rogers, J. Shrinkwrap: Efficient SQL query processing in differentially private data federations. In *Proceedings of the VLDB Endowment*, 2018.
8. Bittau, A. et al. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 2017 Symp. on Operating Systems Principles*.
9. Chaum, D.L. Untraceable electronic mail, return addresses, and digital pseudonyms. *Commun. ACM 24*, 2 (Feb. 1981), 84–90.
10. Cheu, A., Smith, D., Ullman, J., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. *Theory and Practice of Differential Privacy*, 2018.
11. Ding, B., Kulkarni, J., and Yekhanin, S. Collecting telemetry data privately. In *Proceedings of the 2017 Annual Conf. on Neural Information Processing Systems*.
12. Dinur, I. and Nissim, K. Revealing information while preserving privacy. In *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems*. ACM, 2003.
13. Dwork, C. Differential privacy. *Automata, Languages and Programming*. Springer, 2006.
14. Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conf.* Springer, 2006, 265–284.
15. Dwork, C. A. Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 2014.
16. Erlingsson, U., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Annual ACM-SIAM Symp. Discrete Algorithms*, 2019, 2468–2479.
17. Erlingsson, U., Pihur, V., and Korolova, A. RAPPOR: Randomized aggregatable privacy preserving ordinal response. In *Proceedings of the ACM Conf. on Computer and Communications Security*, 2014.
18. Facebook Privacy-Protected URLs Light Table Release; <https://bit.ly/3kKSqXY>.
19. Garfinkel, S., Abowd, J., and Martindale, C.

- Understanding database reconstruction attacks on public data. *Commun. ACM 62* (2019), 46–53.
20. Ghazi, B., Manurangsi, P., Pagh, R., and Velingker, A. Private aggregation from fewer anonymous messages. In *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2020, 798–827.
  21. Groce, A., Rindal, P., and Rosulek, M. Cheaper private set intersection via differentially private leakage. *Privacy Enhancing Technologies Symposium*, 2019.
  22. Johnson, N., Near, J., and Song, D. Towards practical differential privacy for SQL queries. In *Proceedings of the VLDB Endowment*, 2018.
  23. Kasiviswanathan, S., Lee, H., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM J. Comput.* 40, 3 (June 2011), 793–826.
  24. Kasiviswanathan, S., Nissim, K., Raskhodnikova, S., and Smith, A. Analyzing graphs with node differential privacy. *Theory of Cryptography*. Springer, 2013, 457–476.
  25. Kasiviswanathan, S. and Smith, A. On the ‘semantics’ of differential privacy: A Bayesian Formulation. *J. Privacy and Confidentiality*, 2014.
  26. Kifer, D. and Machanavajhala, A. Pufferfish: A framework for mathematical privacy definitions. In *Proceedings of the ACM Trans. Database Systems 39*, 1 (2014).
  27. Lazar, D., Gilad, Y., and Zeldovich, N. Karaoke: Distributed private messaging immune to passive traffic analysis. In *USENIX Symp. on Operating Systems Design and Implementation*, 2018.
  28. Liu, C., Chakraborty, S., and Mittal, P. Dependence makes you vulnerable: Differential privacy under dependent tuples. In *Proceedings of the 2016 Symp. on Network and Distributed System Security*.
  29. Machanavajhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE Intern. Conf. on Data Engineering*.
  30. McGregor, A., Mironov, I., Pitassi, T., Reingold, O., Talwar, K., and Vadhan, S. The limits of two-party differential privacy. In *Proceedings of the 2010 Symp. on Foundations of Computer Science*. IEEE.
  31. McKenna, R., Miklau, G., Hay, M., and Machanavajhala, A. Optimizing error of high-dimensional statistical queries under differential privacy. In *Proceedings of the VLDB Endowment 11*, 10 (2018), 1206–1219.
  32. Mironov, I., Pandey, O., Reingold, O., and Vadhan, S. Computational differential privacy. In *Proceedings of the 29th Annual Intern. Cryptology Conf. Advances in Cryptology*. Springer-Verlag, Berlin, Heidelberg, 2009, 216–242.
  33. Narayan, A. and Haebertlen, A. DJoin: Differentially private join queries over distributed databases. In *Proceedings of the 2012 USENIX Symp. on Operating Systems Design and Implementation*.
  34. Sun, Y., Edmundson, A., Vanbever, L., Li, O., Rexford, J., Chiang, M., and Mittal, P. Raptor: Routing attacks on privacy in Tor. In *Proceedings of the 2015 USENIX Security Symp.*
  35. Tyagi, N., Gilad, Y., Leung, D., Zaharia, M., and Zeldovich, N. Stadium: A distributed metadata-private messaging system. In *Proceedings of the 2017 Symp. on Operating Systems Principles*.
  36. Uber Releases Open Source Project for Differential Privacy; <https://bit.ly/2RV16hX>.
  37. van den Hooff, J., Lazar, D., Zaharia, M., and Zeldovich, N. Vuvuzela: Scalable private messaging resistant to traffic analysis. In *Proceedings of the 2015 Symp. on Operating Systems Principles*.
  38. Wagh, S., Cuff, P. and Mittal, P. Differentially private oblivious ram. In *Proceedings on Privacy Enhancing Technologies 4* (2018, 64–84).
  39. Wolinsky, D., Corrigan-Gibbs, H., Ford, B., and Johnson, A. Dissent in numbers: Making strong anonymity scale. In *Proceedings of the 2012 USENIX Symp. on Operating Systems Design and Implementation*.

**Sameer Wagh** (snwagh@gmail.com) is a post-doc researcher at the University of California, Berkeley, CA, USA.

**Xi He** is an assistant professor in the Cheriton School of Computer Science at the University of Waterloo, Ontario, Canada.

**Ashwin Machanavajhala** is an associate professor and director of Graduate Studies in the Department of Computer Science at Duke University, Durham, NC, USA.

**Prateek Mittal** is an associate professor in the Department of Electrical Engineering at Princeton University, Princeton, NJ, USA.

# Attention: Undergraduate and Graduate Computing Students

There's an **ACM Student Research Competition (SRC)**  
at a SIG Conference of interest to you!



Association for Computing Machinery  
Advancing Computing as a Science & Profession

SPONSORED BY Microsoft

It's hard to put the **ACM Student Research Competition** experience into words, but we'll try...



"Attending ACM SRC was a transformative experience for me. It was an opportunity to take my research to a new level, beyond the network of my home university. Most important, it was a chance to make new connections and encounter new ideas that had a lasting impact on my academic life. I can't recommend ACM SRC enough to any student who is looking to expand the horizons of their research endeavors."

**David Mueller**  
North Carolina State University | SIGDOC 2018



"Participating in the ACM SRC was a unique opportunity for practicing my presentation skills, getting feedback on my work, and networking with both leading researchers and fellow SRC participants. Winning the competition was a great honor, a motivation to continue working in research, and a useful boost for my career. I highly recommend any aspiring student researcher to participate in the SRC."

**Manuel Rigger**  
Johannes Kepler University Linz, Austria | Programming 2018



"The SRC was a great chance to present early results of my work to an international audience. Especially the feedback during the poster session helped me to steer my work in the right direction and gave me a huge motivation boost. Together with the connections and friendships I made, I found the SRC to be a positive experience."

**Matthias Springer**  
Tokyo Institute of Technology | SPLASH 2018



"I have been a part of many conferences before both as an author and as a volunteer but I found SRC to be an incredible conference experience. It gave me the opportunity to have the most immersive experience, improving my skills as a presenter, researcher, and scientist. Over the several phases of ACM SRC, I had the opportunity to present my work both formally (as a research talk and research paper) and informally (in poster or demonstration session). Having talked to a diverse range of researchers, I believe my work has much broader visibility now and I was able to get deep insights and feedback on my future projects. ACM SRC played a critical role in facilitating my research, giving me the most productive conference experience."

**Muhammad Ali Gulzar**  
University of California, Los Angeles | ICSE 2018



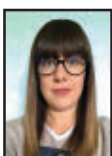
"At the ACM SRC, I got to learn about the work done in a variety of different research areas and experience the energy and enthusiasm of everyone involved. I was extremely inspired by my fellow competitors and was happy to discover better ways of explaining my own work to others. I would like to specifically encourage undergraduate students to not hesitate and apply! Thank you to all those who make this competition possible for students like me."

**Elizaveta Tremsina**  
UC Berkeley | TAPIA 2018



"The ACM SRC was an incredible opportunity for me to present my research to a wide audience of experts. I received invaluable, supportive feedback about my research and presentation style, and I am sure that the lessons I learned from the experience will stay with me for the rest of my career as a researcher. Participating in the SRC has also made me feel much more comfortable speaking to other researchers in my field, both about my work as well as projects I am not involved in. I would strongly recommend students interested in research to apply to an ACM SRC—there's really no reason not to!"

**Justin Lubin**  
University of Chicago | SPLASH 2018



"Joining the Student Research Competition of ACM gave me the opportunity to measure my skills as a researcher and to carry out a preliminary study by myself. Moreover, I believe that "healthy competition" is always challenging in order to improve yourself. I suggest that every Ph.D. student try this experience."

**Gemma Catolino**  
University of Salerno | MobileSoft 2018

Check the SRC Submission Dates: <https://src.acm.org/submissions>

- ◆ Participants receive: \$500 (USD) travel expenses
- ◆ All Winners receive a medal and monetary award. First place winners advance to the SRC Grand Finals
- ◆ Grand Finals Winners receive a handsome certificate and monetary award at the ACM Awards Banquet

**Questions?** Contact Nanette Hernandez, ACM's SRC Coordinator: [hernandez@hq.acm.org](mailto:hernandez@hq.acm.org)





---

P. 96

**Technical  
Perspective  
Programming  
Microfluidics to  
Execute Biological  
Protocols**

By Nada Amin

P. 97

**BioScript: Programming  
Safe Chemistry on  
Laboratories-on-a-Chip**

By Jason Ott, Tyson Loveless, Chris Curtis,  
Mohsen Lesani, and Philip Brisk

---

P. 105

**Technical  
Perspective  
Solving the Signal  
Reconstruction  
Problem at Scale**

By Zachary G. Ives

P. 106

**Scalable Signal  
Reconstruction  
for a Broad Range  
of Applications**

By Abolfazl Asudeh, Jees Augustine, Saravanan Thirumuruganathan,  
Azade Nazi, Nan Zhang, Gautam Das, and Divesh Srivastava

# Technical Perspective

## Programming Microfluidics to Execute Biological Protocols

By Nada Amin

REPRODUCIBILITY OF EXPERIMENTAL results is a cornerstone of biology research. Today, many of these experiments are done using automated machines such as robots and microfluidic chips. However, published reports about the work explain the experimentation method in plain English, which must be interpreted by other groups to reproduce the experiment.

Biological protocols give a recipe for a biological experiment. Ideally, we would like these protocols to be specified rigorously and precisely. Once we do that, we are a step away from automation, reproducibility, and also repurposing.

Microfluidics are diverse technologies to conduct precise and repeatable experiments on small quantities of fluids. Just like integrated circuits have allowed automation of computation, microfluidic chips—coin-sized media that manipulate small quantities of liquid—promise to automate biological and chemical experiments. A common application is DNA replication, enabling small amounts of DNA to be amplified for larger-scale analyses. Reagents are held in chambers on the chip and the interconnecting fluid pathways dictate how and in what proportions reagents are to be mixed; this can be accomplished differently across chip technologies, for example, through microchannels etched into the medium or through electric fields that manipulate discrete droplets. The key benefit of a microfluidic chip is that precise analyses can be performed despite their incredibly small size; this enables massively parallel experiments, low reagent consumption, and overall lower experiment setup costs. Since fabrication of microfluidic devices is cheap, experimenters tend to iterate through their designs quickly. However, the production of lab-on-chips is not purely a matter of manufacturing—bottlenecks exist throughout the microfluidic chip design workflow, including microfluidics-aware computer-aided design tools, design verification, and barriers to entry.


With the advent of microfluidics, there is now a constrained medium in which to explore executable biological protocols. BioStream and BioCoder are the first programming languages for biological protocols in the wake of microfluidics; both are embedded in C++. BioStream separates the specification of the protocol from its realization on a microfluidic chip. BioCoder focuses on expressing high-level protocols and encompasses a variety of biological experiments, leaving the realization of those protocols on microfluidic chips or other media as future work. Developed more recently, BioScript is a simpler stand-alone language with an operational semantics and type system. The type system is based on a table of real hazards and can statically guarantee the experiment does not cause hazards, for example, by mixing incompatible fluids. Even more recently, Puddle, an automation platform based on microfluidics, relies on dynamic feedback rather than static checks to run experiments from bio-computing to medical diagnostics.

The hazard-free guarantee approach taken in the following paper is an example of how programming languages can help develop executable protocols that are conforming, understandable, safe, and retargetable. As in software reuse, one might be able to ‘tweak’ an executable protocol to a new purpose, and one should ensure the guarantees still carry.

Within the next decades, we can imagine that medicine, biology, and chemistry papers that contain results from wet lab experiments come with their own ‘protocol’ artifact and that such artifacts will be more formally specified. It will also be possible to formally analyze the protocols and the results of a paper, to evaluate the claims, and beyond, to evaluate the protocols’ safety, retargetability (running on different hardware), modularity (plugging multiple protocols) and repurposability (running a variant of the protocol).

In another direction, having a programming language to specify a biology protocol that is one step removed from the medium would enable a proliferation of digital media (as in Lab-on-Chips) and compilers-to-chips for more custom media (as in custom chips). We can see the trend that microfluidic chips are either general digital chips that can capture a range of experiments, or custom high-throughput chips that target a particular experiment. From a biology protocol, a compiler could automatically propose how to run the protocol on a prefabricated digital chip or how to fabricate a custom high-throughput chip layout to run the protocol.

Programming microfluidics to execute biological protocols remains an exciting avenue, with a promise still to be fulfilled. Ideally, one should be able to run the same biological protocol code on a variety of potential platforms. Works in programming languages for biological protocols can ensure a separation of concerns between the specification of a biological protocol and its realization on biological media such as microfluidic chips, fostering advances on both sides of the separation.

Finally, creating a faster loop from medical problem to diagnostic to informed decision is a robot scientist’s dream. Programming microfluidics could be put to use at various levels of safety when filtering through the myriad of potential cures to the one cure that will work for the here and now. For example, people are mining SARS-CoV-2-Human Protein-Protein Interaction for drug repurposing. Though robots are less used in high facility labs for obvious ‘gone rogue’ reasons, given the myriad of potential experiments, it would be wonderful to have a tight feedback between the future robot scientist and the robot or human experimenters. 

Nada Amin is an assistant professor of computer science at Harvard SEAS, Cambridge, MA, USA.

Copyright held by author.

# BioScript: Programming Safe Chemistry on Laboratories-on-a-Chip

By Jason Ott, Tyson Loveless, Chris Curtis, Mohsen Lesani, and Philip Brisk

## Abstract

This paper introduces *BioScript*, a domain-specific language (DSL) for programmable biochemistry that executes on emerging microfluidic platforms. The goal of this research is to provide a simple, intuitive, and type-safe DSL that is accessible to life science practitioners. The novel feature of the language is its syntax, which aims to optimize human readability; the technical contribution of the paper is the *BioScript* type system. The type system ensures that certain types of errors, specific to biochemistry, do not occur, such as the interaction of chemicals that may be unsafe. Results are obtained using a custom-built compiler that implements the *BioScript* language and type system.

## 1. INTRODUCTION

The last two decades have witnessed the emergence of software-programmable laboratory-on-a-chip (pLOC) technology, enabled by technological advances in microfabrication and coupled with scientific understanding of microfluidics, the fundamental science of fluid behavior at the micro- to nanoliter scale. The net result of these collective advancements is that many experimental laboratory procedures have been miniaturized, accelerated, and automated, similar in principle to how the world's earliest computers automated tedious mathematical calculations that were previously performed by hand. Although the vast majority of microfluidic devices are effectively application-specific integrated circuits (ASICs), a variety of programmable LoCs have been demonstrated.<sup>16,18</sup>

With a handful of exceptions, research on programming languages and compiler design for programmable LoCs has lagged behind their silicon counterparts. To address this need, this paper presents a domain-specific programming language (DSL) and type system for a specific class of pLOC that manipulate discrete droplets of liquid on a two-dimensional grid. The basic principles of the language and type system readily generalize to programmable LoCs, realized across a wide variety of microfluidic technologies.

The presented language, *BioScript*, offers a user-friendly syntax that reads like a cookbook recipe. *BioScript* features a combination of fluidic/chemical variables and operations that can be interleaved seamlessly with computation, if desired. Its intended user base is not traditional software developers, but life science practitioners, who are likely to balk at a language that has a steep learning curve.

*BioScript*'s type system ensures that each fluid is never consumed more than once, and that unsafe combinations

of chemicals—those belonging to conflicting reactivity groups, as determined by appropriately qualified government agencies—never interact on-chip; *BioScript*'s type system is based on union types and was designed to ensure that type inference is decidable. This will set the stage for future research on formal validation of biochemical programs.

The *BioScript* language and type system are evaluated using a set of benchmark applications obtained from scientific literature. We use a microfluidic simulator to assess performance under ideal operating conditions and also execute them on a real device, which is much smaller and supports a subset of *BioScript*'s operational capabilities. This result establishes the feasibility of high-level programming language and compiler design for programmable chemistry, and opens up future avenues for research in type systems and formal verification techniques within this nontraditional computing domain.

### 1.1. Digital microfluidic biochips (DMFBs)

This paper targets a specific class of programmable LoCs that manipulate discrete droplets of fluid via electrostatic actuation. Figure 1a illustrates the electrowetting principle: applying an electrostatic potential to a droplet modifies the shape of the droplet and its contact angle with the surface.<sup>10,13</sup> As shown in Figure 1b, droplet transport can be induced by activating and deactivating a sequence of electrodes adjacent to the droplet; the ground electrode, on top of the array, improves the fidelity of droplet motion and reduces the voltage required to induce droplet transport.

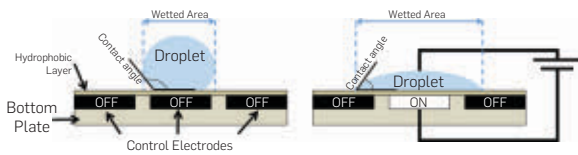
Figure 2a depicts a programmable 2D electrowetting array, called a digital microfluidic biochip (DMFB). A DMFB can support five basic operations, as shown in Figure 2b: transport (move a droplet from position  $(x, y)$  to  $(x', y')$ ), split (create two smaller droplets from one larger droplet), merge (combine two droplets into 1), mix (rotate a merged droplet in a rectangular region around one or more pivots), and storage (place a droplet at position  $(x, y)$  for later use). A DMFB is reconfigurable, as these operations can be performed anywhere on the array, and any given electrode can be used to perform different operations at different times. Droplet I/O is performed using reservoirs on the perimeter of the chip, which are not depicted in Figure 2.

The DMFB instruction-set architecture (ISA) can be extended by integrating sensors, optical detectors, or

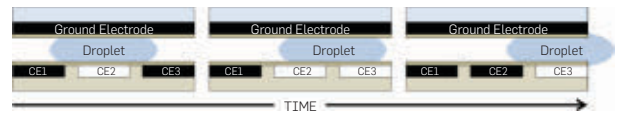
The original version of this paper was published in the *Proceedings of OOPSLA '18* (Boston, MA, Nov. 2018), Article 128.



**Figure 1. The electrowetting principle (a) enables droplet transport (b).**

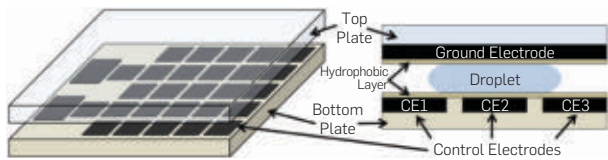


(a) The electrowetting principle:<sup>10, 13</sup> applying an electrostatic potential to a droplet at rest reduces the contact angle with the surface, thereby increasing the surface area in contact with the droplet



(b) A droplet is transported from control electrode CE2 to neighboring electrode CE3 by activating CE3, and then deactivating CE2 (white: activated electrode; black: deactivated electrode).

**Figure 2. A DMFB (a) and its reconfigurable instruction set (b).**



(a) Left: A DMFB is a planar array of electrodes.<sup>15</sup> Right: Cross-sectional view.



(b) The DMFB ISA supports five basic operations: transporting, merging, splitting, mixing and storage, in addition to I/O on the perimeter of the array.

online video monitoring capabilities. Sensors and actuators create a “cyber-physical” feedback loop between the host PC controller and the DMFB. The ability to perform sensing, computation, and actuation based on the results of the computation adds control flow to the instruction set of the DMFB. Prior work has applied feedback control for precise droplet positioning and online error detection and recovery<sup>11, 12, 19</sup> efforts to leverage these capabilities to provide control flow constructs at the language syntax level have been far more limited.

## 2. OVERVIEW

**BioScript Syntax and Semantics.** *BioScript* is a language for programmable microfluidics whose syntax aims to be palatable to life science practitioners, most of whom are not experienced programmers. The *BioScript* syntax and semantics were designed to enable scientists to express operations in a manner that closely resembles plain English. To keep the language small, we do not include operations in the language syntax that can automatically be inferred by the compiler and/or execution engine. For example, the compiler can automatically infer implicit fluid transfers for a mix operation. *BioScript* features a semantics that targets pLOC technologies. The syntax and semantics of *BioScript*'s type system are formally described in Section 3.

We begin with a self-contained example to illustrate the expressive capabilities of *BioScript*.

**Example: PCR with Droplet Replenishment.** Figure 3 presents a *BioScript* program for a DMFB-compatible implementation of the *polymerase chain reaction (PCR)*, used to amplify DNA.<sup>14</sup> PCR involves *thermocycling* (repeatedly heating and then cooling) a droplet containing the DNA mixture undergoing amplification [lines 5–17]. In this implementation, thermocycling may cause excess droplet evaporation. This implementation uses a weight sensor to detect the droplet

**Figure 3. PCR with droplet replenishment.<sup>9</sup> It uses the target-specific save instruction.**

```

1 // Initialization omitted. PCRMasterMix is a
2 // commercially available pre-mixed solution
3 // used to perform PCR.
4 PCRMix = mix PCRMasterMix with Template for 1s
5 repeat 50 times {
6   heat PCRMix at 95C for 20s
7   volumeWeight = detect Weight on PCRMix
8   if (volumeWeight <= 50uL) {
9     replacement = mix 25uL of PCRMasterMix
10      with 25uL of Template for 5s
11     heat replacement at 95C for 45s
12     PCRMix = mix PCRMix with replacement for 5s
13   }
14   heat PCRMix at 68C for 30s
15   heat PCRMix at 95C for 45s
16 }
17 heat PCRMix at 68C for 5min
18 save PCRMix

```

volume after each iteration [line 8]; if too much evaporation occurs [line 9], the algorithm injects a new droplet to replenish the sample volume [line 10–11], preheating a template solution [line 12] to ensure that replenishment does not affect the temperature of the DNA.

**Type Systems and Safety.** The Environmental Protection Agency (EPA) and National Oceanic and Atmospheric Administration (NOAA) have categorized 9800 chemicals into 68 reactivity groups,<sup>7</sup> defined by common physical properties of discrete chemicals. It is known that mixing materials from certain reactivity groups can produce materials from other reactivity groups; for example, mixing acids and bases induces a strong reaction that produces salt and water. *BioScript*'s type system models reactivity groups as types. As a material can belong to multiple reactivity groups, a union type is associated with a material. Using standard reaction corpora, we calculate the type signature of the mix operation, which is fundamental throughout chemistry, as

a table of abstract reactions between pairs of types, which results in a union of types.

At the same time, reactions vary in terms of safety. The EPA/NOAA categorization assigns one of three outcomes to the combination of chemicals: *Incompatible*, *Caution*, or *Compatible*. If the union type resulting from a mix operation includes a hazardous type, then the corresponding cell in the table is marked as being unsafe. Any biochemical procedure, or *assay*, specified in *BioScript* is allowed to execute only if it is safe. The signature of the mix operation does not include unsafe abstract reactions, which correspond to unsafe table cells. Therefore, the type system exclusively type-checks mix statements that do not produce hazardous materials. This is fundamental to the soundness of *BioScript*'s type system: it only type-checks assays that do not produce unsafe materials.

*BioScript* allows, but does not require, type annotations, saving the programmer from the burden of annotating programs with overly complicated union types. The assay specifications presented in Figure 3 do not use type annotations. *BioScript*'s type inference system can automatically infer types. As the EPA/NOAA classification begins with a finite set of material types, type inference can be reduced to efficiently decidable theories. Type inference is sound: if a typing assignment is inferred, it can be used to type-check the assay; it is also complete: if there is a typing assignment with which the assay can be type-checked, the inference will discover it. Otherwise, the assay is rejected and marked as a potential hazard if no typing assignment can be inferred for it. Our experiments show that the type system is expressive enough to reject hazardous assays and accept those that are safe. Proofs for these attributes can be found in Ott et al.<sup>15</sup> and its supplemental material.

### 3. TYPE SYSTEM

This section presents interesting aspects of the core *BioScript* language. We begin by presenting the simple, yet robust, *BioScript* syntax. Next, we describe the novel aspects of the operational semantics—or mathematical model—describing the runtime execution of assays on pLoC devices (Definition 1). We then provide technical details on how *BioScript*'s type system prevents unsafe operations from occurring. Unsafe operations include the interaction of materials that may cause an explosion or create noxious gasses, as well as access materials that have already been consumed. We explore the syntax, operational semantics, and type system using two statements: variable assignment and mix semantics in great detail. The full *BioScript* language, operational semantics, and type system are described in Ott et al.<sup>15</sup> and its supplemental material.

#### 3.1. Syntax

*BioScript*'s set of instructions is modeled after the ISA discussed in Section 1.1. *BioScript* supports heat and detect instructions but omits move and store instructions, as they are inferred from data-flow analysis. The *BioScript* language is imperative and a statement is a sequence of effectful instructions that involve side effect-free terms. To model state, or memory, we define  $\sigma$ , a mapping of variables to their values. A side effect, in this context, is changing  $\sigma$ —updating

the values of the variables. As terms are side effect-free, a term does not alter  $\sigma$ . A term can take one of many forms: a variable, a math operation, a detection of a physical property on a material, or a concrete value.

Unlike terms, instructions are not side effect-free; they alter memory. *BioScript* supports traditional assignment of terms to variables, manipulation of variables, and control-flow constructs.

*BioScript* utilizes a conservative type system capable of analyzing how chemical interactions work in a cyber-physical context. Mixing chemicals during experiments yields a new chemical, functionally expiring the input chemicals. However, not all of the input chemicals participate in the reaction, and trace amounts of the input chemicals are present in the new chemical. For instance, mixing an acid and base yields salt water. There are still acid and base molecules that have not reacted in the salt water. To model this, *BioScript* employs union types that allow variables to belong to multiple types (see Definition 2). In other words, a variable can store any combination of scalar types in the union type. As usual, the typing environment  $\Gamma$  represents a mapping from variables to their types.

#### Operational Semantics:

The operational semantics describe how a program is executed as a sequence of computational steps. It is represented as inference rules that define valid steps. Inference rules are comprised of premises and conclusions, whereby all the premises must be met for the conclusion to hold. As shown in Figures 4 and 5, the inference rules represent the premises above the line and conclusion below the line.

#### Definition 1

#### 3.2. Operational Semantics for Assay Execution

We model execution of *BioScript* assays on a DMFB as an operational semantics. When execution of an instruction occurs, for example, a mix, the model must use the appropriate rules to “step” or handle the change of state. All the premises of a stepping rule must be satisfied. If no rule can step, the program is *stuck* and cannot continue execution.

We highlight two sets of rules that showcase some interesting challenges *BioScript* faces and discuss how they are overcome. We begin with variable assignment. It is syntax:  $x := t$  allows a variable  $x$  to be assigned some term  $t$ . To model execution, we define E-ASSIGNR, E-ASSIGN, and E-ASSIGN', represented in Figure 4a.

The rule E-ASSIGNR evaluates the right-hand side term,  $t$  (if it is not a variable); the rule E-ASSIGN assigns the reduced value to the variable in  $\sigma$ , the store. The rule E-ASSIGN' transfers a material from the right-hand side variable to the left-hand side variable; preventing aliasing.

In traditional computing, variable assignment is an elementary operation that most computer scientists do not even bother thinking about. However, when modeling assignment in the physical world, things are not so simple. In *BioScript*, the value of a chemical variable is consumed when it is assigned to another variable, restricting variable aliasing. In other words,

**Figure 4. a and b depict the operational semantics for only variable assignment and mixing in *BioScript*.**

$$\frac{\text{E-ASSIGNR} \quad (\sigma, t) \rightarrow t' \quad t \notin \mathcal{X}}{(\sigma, x := t; s) \rightarrow (\sigma, x := t'; s)} \quad \text{E-ASSIGN} \quad (\sigma, x := v; s) \rightarrow (\sigma[x \mapsto v], s)$$

$$\frac{\text{E-ASSIGN}' \quad \sigma' = (\sigma \setminus \{x'\})[x \rightarrow \sigma(x')]}{(\sigma, x := x'; s) \rightarrow (\sigma', s)} \quad \text{(a)}$$

$$\frac{\text{E-MIXR} \quad (\sigma, t) \rightarrow t'}{(\sigma, x := \mathbf{mix} \ x_1 \ \mathbf{with} \ x_2 \ \mathbf{for} \ t; s) \rightarrow (\sigma, x := \mathbf{mix} \ x_1 \ \mathbf{with} \ x_2 \ \mathbf{for} \ t'; s)}$$

$$\frac{\text{E-MIX} \quad \begin{array}{l} \sigma(x_1) \in \text{Mat} \quad \sigma(x_2) \in \text{Mat} \\ \text{interact}(\sigma(x_1), \sigma(x_2), r) \neq \perp \\ \sigma' = (\sigma \setminus \{x_1, x_2\})[x \mapsto \text{interact}(\sigma(x_1), \sigma(x_2), r)] \end{array}}{(\sigma, x := \mathbf{mix} \ x_1 \ \mathbf{with} \ x_2 \ \mathbf{for} \ r; s) \rightarrow (\sigma', s)} \quad \text{(b)}$$

the *BioScript* program  $x = \text{mat}; y = x; z = x$  is stuck at the third assignment as the second assignment consumes  $x$ . This restriction is necessary for material variables, but can be easily lifted for numeric variables.

Mixing is a frequent activity that chemists and biologists employ in their discipline. *BioScript*'s syntax for mixing is simple and intuitive:  $x := \text{mix } x_1 \ \text{with } x_2 \ \text{for } t$ . A mix instruction takes two variables ( $x_1$  and  $x_2$ ), mixes them for some time  $t$  and stores the resulting chemical in the new variable  $x$ . To model execution of the mix instruction, we define E-MIXR and E-MIX, defined in Figure 4b.

E-MIXR first evaluates the time term of a mix instruction, eventually reducing it to a real number,  $r$ . After the time term has been reduced, E-MIX is evaluated. E-MIX prescribes that both  $x_1$  and  $x_2$  in  $\sigma$  must be materials. The variables  $x_1$  and  $x_2$  must also be safe to interact; the function *interact* determines safety at run time. *interact* returns the resulting material if mixing is safe; otherwise, *interact* returns  $\perp$ —the mixture is unsafe. When a scientist mixes two chemicals together in a flask, the two distinct chemicals no longer exist; to model this, the used variables  $x_1$  and  $x_2$  are removed from  $\sigma$  and the variable  $x$  is mapped to the resulting material. The evaluation of a mix instruction is *stuck* if either of the two variables are not material values, any of the variables are already used and removed from the store, or the interaction of the materials is unsafe ( $\perp$ ).

The full runtime model, detailing all terms and instructions, is available in the supplemental material.

### 3.3. Type Checking and Inference

Similar to modeling execution, inference rules describe how *BioScript*'s type system type-checks a program. Again, we focus on the interesting typing rules that *BioScript* defines to keep scientists safe while writing and executing assays on DMFB devices.

**Figure 5. a and b depict the typing rules for only variable assignment and mixing in *BioScript*.**

$$\frac{\text{T-ASSIGN-1} \quad x: T \in \Gamma \quad \Gamma, X \vdash v: T' \quad T' \subseteq T}{\Gamma, X \vdash x := v, X \cup \{x\}}$$

$$\frac{\text{T-ASSIGN-2} \quad x: T \in \Gamma \quad \Gamma, X \vdash x': T' \quad T' \subseteq T}{\Gamma, X \vdash x := x', X \setminus \{x'\} \cup \{x\}}$$

$$\frac{\text{T-ASSIGN-3} \quad \begin{array}{l} x: T \in \Gamma \quad t \notin \mathcal{V} \cup \mathcal{X} \\ \Gamma, X \vdash t: T' \quad T' = \mathbb{R} \vee T' = \mathbb{N} \quad T' \subseteq T \end{array}}{\Gamma, X \vdash x := t, X \cup \{x\}}$$

$$\frac{\text{T-MIX} \quad \begin{array}{l} \Gamma, X \vdash x_1: \cup \overline{\text{Mat}}_i \quad \Gamma, X \vdash x_2: \cup \overline{\text{Mat}}_j \quad \Gamma, X \vdash t: \mathbb{R} \\ \text{interact-abs}(\text{Mat}_i, \text{Mat}_j) \subseteq \Gamma(x) \text{ for each } i \text{ and } j \end{array}}{\Gamma, X \vdash x := \mathbf{mix} \ x_1 \ \mathbf{with} \ x_2 \ \mathbf{for} \ t, X \setminus \{x_1, x_2\} \cup \{x\}}$$

We begin with typing assignment instructions, defined in Figure 5a. The rule T-ASSIGN-1 types an assignment of a value to a variable and adds the variable to the set of available variables. Rule T-ASSIGN-2 strictly prevents aliasing by consuming the right-hand side while adding the left-hand side variable to the set of available variables. (At the cost of brevity, the rule can be easily relaxed to not remove numeric variables from the available set.) Finally, rule T-ASSIGN-3 addresses typing for numeric terms. It allows assigning numeric terms to variables.

In spite of a scientist's training regarding safe and unsafe chemical interactions, countless incidents occur involving chemical interactions that result in explosions or noxious gasses, causing harm to the laboratory or worse, the scientist. To help prevent incidents, *BioScript* defines the typing rule T-MIX, described in Figure 5b, which helps ensure that no chemical interaction exhibits adverse reactions as well as guaranteeing no chemical is used more than once.

To guarantee safety during a mix instruction,  $x_1$  and  $x_2$  must be a union of material types, that is,  $\Gamma, X \vdash x_1: \cup \overline{\text{Mat}}_i$  and  $\Gamma, X \vdash x_2: \cup \overline{\text{Mat}}_j$ , respectively. Similarly, the time term of the mix instruction must be a real number ( $\Gamma, X \vdash t: \mathbb{R}$ , which is to say that the value of the term  $t$  must be in the set of real numbers).

#### Union Types:

A typing convention allows a variable to assume a set of types. We differentiate between *scalar types*, denoted by  $S$ , and *union types*, denoted by  $\cup \overline{S}$ ; a union type is a set of scalar types. In the context of *BioScript*, scalar types are the material types  $\text{Mat}_1 | \dots | \text{Mat}_n$ . A union of material types can then be expressed as  $\cup \overline{\text{Mat}}$ .

#### Definition 2



For a mix instruction to type-check, the interaction of the input materials must be safe. To determine this, we define the function *interact-abs*, which accepts two scalar material types as arguments and returns a union type of materials ( $\cup \text{Mat}$ ). The abstract interaction *interact-abs* is conservative with respect to the concrete interaction function: *interact*. If two material values  $mat_i$  and  $mat_j$  are members of two material types  $Mat_i$  and  $Mat_j$ , and the concrete interaction of  $mat_i$  and  $mat_j$  is unsafe, then the abstract interaction of  $Mat_i$  and  $Mat_j$  is undefined, rendering the program unable to type-check. Otherwise, the result of the concrete interaction is a member of the type resulting from the abstract interaction of  $Mat_i$  and  $Mat_j$ . If the interaction of all such pairs of materials  $mat_i$  and  $mat_j$  is safe, then the abstract interaction of  $Mat_i$  and  $Mat_j$  is safe. A full discussion of how the *interact-abs* function is used is presented in Section 4.

Finally, the result of the mix is assigned to  $x$ , whose type in  $\Gamma$  should be a superset of the resulting material types. In the physical world, mixing chemicals uses those chemicals—they no longer exist. To model this, the materials represented by  $x_1$  and  $x_2$  are consumed and replaced by  $x$  in the set of available variables.

We proved that the *BioScript* type system is sound. All type-checked programs are correct, that is, never get stuck during execution; conversely, incorrect programs cannot type-check. As explained for the operational semantics, there is no inference rule for unsafe operations; that is, incorrect programs are stuck. The soundness is proved as tandem progress and preservation lemmas (see Definition 3). The progress lemma states that well-typed programs are not stuck; that is, they can take a step. More precisely, if a statement is typed, then it is either the terminal statement or it can make a step. The preservation lemma states that if a well-typed program steps, the resulting program is also well-typed.

*BioScript* features a type inference system. Type inference helps the biologists and chemists by lifting the burden of manually annotating assays with union types. The rules for type inference match the corresponding type-checking rules but restate the conditions as constraints. After the type inference system derives the constraints for a program, a satisfying model for the constraints yields types for the variables of the program. We proved that the type inference system infers types for a program if it is typeable. This is proved as a pair of soundness and completeness lemmas for the type inference system. The soundness lemma states that, if the type inference system infers types for a program, then with the inferred types, the type-checking system can type-check the program. The completeness lemma states that if, for a program, there exist types for variables under which the type-checking system can type-check the program, then the type inference system can infer those types.

We provide a full discussion of the above theorems in the supplemental materials for the interested reader.

**Progress:**

A well-typed program is not stuck: that is, it can take a step.

**Preservation:**

If a well-typed term takes a step, the resulting term is also well-typed.

**Definition 3**

## 4. IMPLEMENTATION

This section describes the underlying implementation details of the *BioScript* language and its type system.

**BioScript.** The *BioScript* language was implemented as described in Section 2. As DMFBs do not offer external fluidic storage, there is no possibility to implement a stack or heap of substantial size. For these reasons, *BioScript* provides *inline* functions exclusively and does not support recursion; similarly, *BioScript* does not support arrays, even of constant size, as doing so would significantly inhibit portability. We hope to address these issues in greater detail in a future publication. *BioScript* handles variable assignment implicitly, for example, Figure 8d. However, the scientist declares a manifest of chemicals that is used throughout the assay (“blood” and “water,” for this assay) and the *BioScript* compiler infers the *dispense* and *move* operations.

**The Type System.** *BioScript*'s type system utilizes static type checking, which runs during compilation. The type system automatically infers types using an abstract interaction function that is a conservative overapproximation of the resulting chemical types of each interaction. The type system uses the 68 EPA/NOAA reactivity groups as the material types  $\text{Mat}_i$ , that together with natural  $\mathbb{N}$ , and real  $\mathbb{R}$  numbers, constitute the set of scalar types  $S$ .

We calculate the abstract interaction function *interact-abs* (defined in Section 3.3) as a table that is indexed by two material types and stores union types. Each reactivity group or type  $Mat_i$  comprises a nonempty set of chemicals  $C_i$ . Abstract mixing of a pair of material types  $Mat_i$  and  $Mat_j$  effectively mixes each pair of chemicals  $(c_i, c_j)$  in the cross product  $C_i \times C_j$ . If any interaction is *Incompatible*, the table entry for  $(Mat_i, Mat_j)$  is marked as hazardous (or undefined, as modeled in Section 3). Otherwise, if the mix operation yields a new chemical  $c_k$ , we use a *ChemAxon*,<sup>4</sup> an industry-standard computational chemistry library to assign a union type  $\cup \text{Mat}_k$  to  $c_k$ , which is added to the union type of the cell for  $Mat_i$  and  $Mat_j$ . In practice, molecules of  $c_i$  and  $c_j$  will remain after mixing  $c_i$  and  $c_j$ , even if a reaction occurs, and the presence of extra molecules at the microliter scale, or smaller, may have a nonnegligible impact on the underlying chemistry or biology. To account for this fact,  $Mat_i$  and  $Mat_j$  are also added to the cell. As type assignment to concrete chemicals is conservative and we include the input types in the resulting union type, the types in the table represent an overapproximation of the chemicals that can result from concrete interactions.

There may be instances where scientists need to create hazardous reactions, which the type system would correctly reject. In this case, the type system generates all relevant errors and warnings, but allows the programmer to override the type system in order to finish compilation and execute the assay.

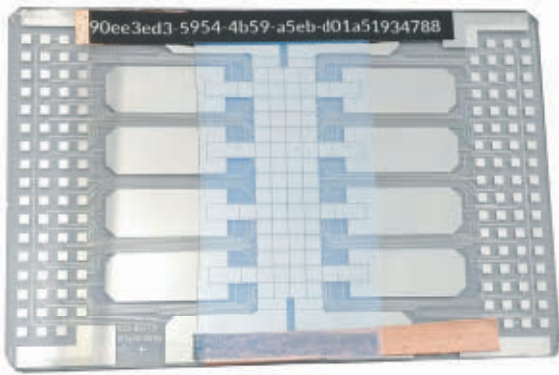
**Execution.** *BioScript* targets a real-world DMFB platform called DropBot,<sup>8</sup> as shown in Figure 6. Although DropBot features real-time object tracking, it does not, at present, support execution of assays that feature control flow. *BioScript* can produce a DropBot-compatible electrode activation sequence, in the form of a JSON file, to execute on the chip depicted in Figure 6.

## 5. EVALUATION

The objectives of *BioScript* are to reduce the time and cost of scientific research and to provide a safe execution environment for chemists and biologists with respect to chemical interactions. As noted earlier, *BioScript* is a DSL that enables high-level programming and direct execution of bioassay on pLoCs. These objectives inform our selection of metrics to evaluate *BioScript*.

**Language.** Compared to other languages, *BioScript* offers an intuitive and readable syntax and a type system. We do not claim that *BioScript* offers any performance advantages over other languages; performance primarily depends on the algorithms implemented in the compiler back-end and execution engine, which are compatible, in principle, with any language and front-end. Hence, our evaluation emphasizes qualitative metrics of the language.

Figure 6. A DMFB chip used by DropBot devices.

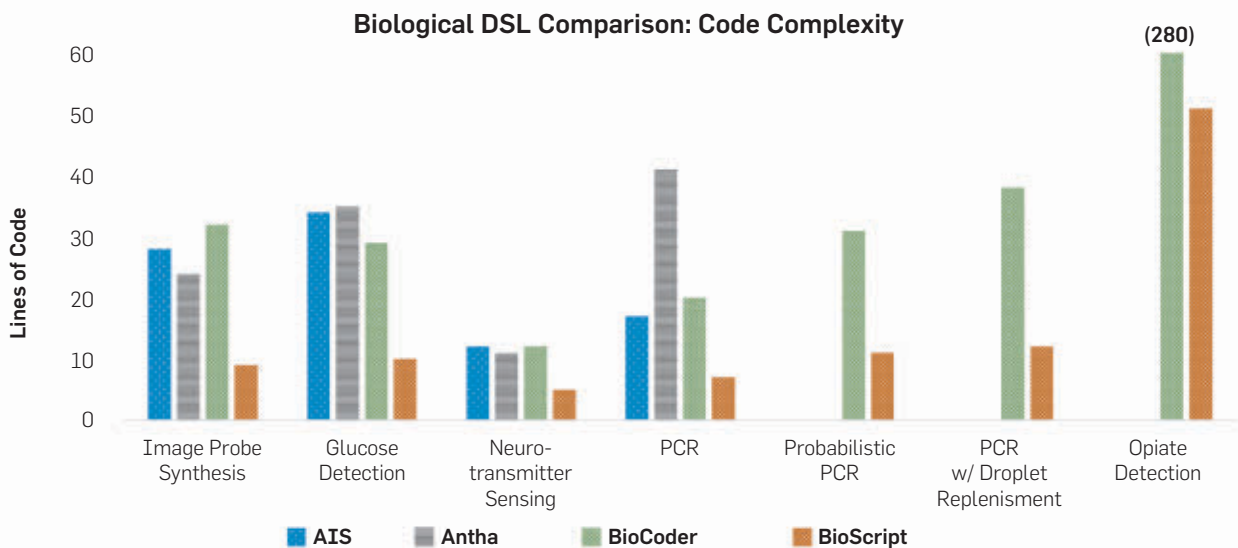


First, we compare *BioScript*'s syntax to three other languages: the *AquaCore Instruction Set (AIS)*, a target-specific assembly-like language;<sup>2</sup> *Antha*, a language for cloud-based laboratory automation;<sup>17</sup> and *BioCoder*, a C++ library that has been previously specialized for DMFBs.<sup>5</sup> Our comparison uses a set of compact, yet representative, bioassays taken from published literature. As an illustrative example, Figure 8 shows a simple assay (a Mix followed by a Heat instruction) in all four languages; *BioScript*, by far, has the shortest description and is easier to read.

Figure 7 compares the number of lines of code required to specify seven representative bioassays using the four languages; three of the seven assays were not compatible with *AIS* (which is tethered to a specific pLoC<sup>2</sup>) and *Antha* (which is tethered to a cloud laboratory), so we only report four assays for those languages. We do not count empty lines (for spacing/aesthetic purposes) or lines that contain comments. We wrote each assay based on our notion of human readability, which generally meant one statement/operation per line for *AIS*, *BioCoder*, and *Antha*. As shown in Figure 8d, the mixture statement in *BioScript* succinctly encompasses two implicit variable declarations with fluid type and volume information.

Across the four compatible assays, *BioScript* required 68% fewer lines of code than *AIS* and 73% fewer lines of code than *Antha*. Across all seven assays, *BioScript* required 65% fewer lines of code than *BioCoder*, which can target DMFBs, unlike *AIS* and *Antha*. Although these results do not account for subjective experience, we believe that they convey the same basic sentiments as shown in Figure 8: *BioScript* has an intuitive syntax and will be far easier for scientists to learn and use compared to existing languages in this space. Source code for all implementations of the bioassays reported in Figure 7 is included in our supplementary materials.

Figure 7. The number of lines of code to specify image probe synthesis, glucose detection, neurotransmitter sensing, PCR, probabilistic PCR, PCR w/ droplet replacement, and opiate detection in *AIS*, *BioCoder*, *Antha*, and *BioScript*. We were unable to specify the latter three assays in *AIS* and *Antha*.



**Type System Evaluation.** *BioScript*'s type system's main purpose is to prevent inadvertent production of hazardous chemicals. We evaluate its ability to detect hazardous mixing in *BioScript* descriptions of five reported real-world incidents.<sup>1,3</sup> To the best of our understanding, *BioScript*'s type system is first-of-its-kind, so there are no prior type systems for biochemistry to compare against.

Table 1 summarizes the results of our experiments. The results denoted by the † are real-world situations in which safety precautions were ignored while carrying out experiments. The first three are incidents documented by the *American Industrial Hygiene Association (AIHA)*.<sup>1</sup> *Mustard gas* refers to a documented situation where an individual mixed two common reagents used to clean swimming pools, inadvertently creating mustard gas. *SafetyZone* refers to a documented explosion where a student mixed a sulfuric acid/hydrogen peroxide mixture with acetone<sup>6</sup> (it remains unknown whether this explosion was intentional or accidental). The type system correctly identified the presence of safety hazards in all of these cases.

We also tested the type system on 14 assays that were known to be safe; *BioScript*'s type system successfully inferred types in all of these cases. These assays, listed in Table 1, are currently used in the physical sciences today.

**Compilation Time.** We compiled the safe and unsafe assays described here, targeting the DropBot platform, which is a 4×15 array (not including I/O reservoirs which reside on the perimeter of the device), assuming the default electrode actuation time of 750 ms. The experiments were

**Figure 8. Example assay specified using BioCoder (a), Antha (b), AIS (c), and BioScript (d). We omit initialization for all examples.**

```
1 b.first_step();
2 b.measure_fluid(blood, tube);
3 b.measure_fluid(water, tube);
4 b.next_step();
5 b.tap(tube, tenSec);
6 b.next_step();
7 b.incubate(tube, 100, tenSec);
8 b.end_protocol();
```

(a)

```
1 smpl := make([]*wtype.LHComponent, 0)
2 Bld := mixer.SampleForTotalVolume(Blood, BldVol)
3 smpl = append(smpl, Bld)
4 Wtr := mixer.Sample(Water, WtrVol)
5 smpl = append(smpl, Wtr)
6 rctn := MixInto(OutPlate, "", smpl...)
7 rl := Incubate(rctn, mltTemp, InitDenatime, false)
```

(b)

```
1 input s1, ip1
2 input s2, ip2
3 move mixer1, s1;
4 move mixer1, s2;
5 mix mixer1, 10;
6 move heater1, mixer1;
7 incubate heater1, 100, 10;
```

(c)

```
1 mixture = mix water with blood for 10s
2 heat mixture at 100C for 10s
```

(d)

run on a 2.7 GHz Intel™ Core i7 processor, 8 GB RAM, machine running macOS™. Construction of the type system's *abstract interaction table* took 31 min running on a 2.53 GHz Intel™ Xeon™ processor, with 24 GB RAM, running CentOS 5.

Table 1 reports the compilation time, constraint solving time, and number of constraints gathered. The unsafe, real-world, assays were correctly identified as unsafe by *BioScript*. On average, each material defined in the benchmarks belonged to 3.015 distinct reactive groups; average benchmark compilation time was 0.0190 s; and the average time spent solving constraints was 1.594 s. We must note that these programs are significantly smaller than typical software programs today.

*BioScript* assays, along with additional synthetic benchmarks, are made available in the supplemental materials.

## 6. CONCLUSION AND FUTURE WORK

*BioScript* enables scientists to express assays in a comfortable manner, similar in principle to laboratory notebooks. Its type system, which defines the operational semantics of *BioScript*, can provide safety guarantees for chemicals used. *BioScript* is extensible, allowing it to target pLoC compilation and LoC synthesis across multiple technologies. *BioScript* and its software stack pave the way for many life science subdisciplines to increase productivity due to automation and programmability. This paper reports a full system implementation, which can compile and type-check a high-level language program and execute it on the real-world DropBot platform by transmitting commands (electrode actuation sequences) via the DropBot software interface.

Being nascent, *BioScript*'s type system statically type-checks only chemical reactivity groups. Extending the type system, introducing dependent types to account for properties such as temperature, pH, volume, or concentration is a natural next step.

**Table 1. Compile time and the number of constraints gathered**

Benchmark	Compile time (s)	Type check time (s)	Total types
AIHA 1†	0.012	0.936	70
AIHA 2†	0.012	1.648	68
AIHA 3†	0.014	1.214	17
Broad spectrum opiate	0.011	0.887	11
Ciprofloxacin	0.023	1.722	14
Diazepam	0.024	1.007	14
Dilution	0.014	0.892	9
Fentanyl	0.018	0.900	13
Full morphine	0.048	4.188	19
Glucose detection	0.012	1.633	14
Heroin	0.020	1.553	13
Image probe synthesis	0.015	2.181	13
Morphine	0.018	1.026	13
Mustard gas†	0.015	1.433	83
Oxycodone	0.026	0.959	13
PCR	0.032	3.534	8
Safety zone†	0.013	1.341	76

†Real-world instances that resulted in damages to equipment or personnel that the type system was correctly able to identify as dangerous.



In the long term, this type system could be generalized into a generic type system for cyber-physical systems, transcending even pLoC-based biochemistry. In the future, we hope to extend the *BioScript* language with support for noninlined functions, arrays, SIMD operations, and some notion akin to processes or threads. We view the type system as a starting point for a much deeper foray into formal verification, for example, to ensure that biological media always experience physical properties such as temperature or pH levels within a user-specified range.

### Acknowledgments

We would like to thank Philipp Haller for his feedback and insight on elements of this work. 

### References

- American Industrial Hygiene Association. 2016. <http://bit.ly/2eZtf1m>. [Accessed: 2016-11-08].
- Amin, A.M., Thottethodi, M., Vijaykumar, T.N., Wereley, S., Jacobson, S.C. Aquacore: A programmable architecture for microfluidics. In D.M. Tullsen, and B. Calder, eds. Proceedings of the 34th International Symposium on Computer Architecture (ISCA 2007), June 9–13, 2007, San Diego, California, USA, ACM, 2007. pp. 254–265
- Blog SPH. Swimming pool chemical incident. 2016. <http://bit.ly/2gghGZI>. [Accessed: 2016-11-01].
- ChemAxon. 2016. <http://www.chemaxon.com>. Marvin was used for characterizing chemical structures, substructures and reactions. Marvin 16.10.3.
- Curtis, C., Brisk, P. Simulation of feedback-driven PCR assays on a 2d electrowetting array using a domain-specific high-level biological programming language. *Microelectronic Engineering* 148, (2015), 110–116.
- Dobbs, D.A., Bergman, R.G., Theopold, K.H. Piranha solution explosion. 1990.
- Environmental Protection Agency & National Oceanic and Atmospheric Administration. 2016. <https://cameochemicals.noaa.gov/>.
- Fobel R, Fobel C, Wheeler AR. Dropbot: An open-source digital microfluidic control system with precise control of electrostatic driving force and instantaneous drop velocity measurement. *Appl. Phys. Lett.* 19, 102 (2013).
- Jebrail, M.J., Renzi, R.F., Sinha, A., Van De Vreugde, J., Gondhalekar, C., Ambriz, C., Meagher, R.J., Branda, S.S. A solvent replenishment solution for managing evaporation of biochemical reactions in air-matrix digital microfluidics devices. *Lab Chip* 15, (2015), 151–158.
- Lippmann, G. Relations entre les phénomènes électriques et capillaires. Gauthier-Villars. 1875.
- Luo, Y., Chakrabarty, K., Ho, T. Error recovery in cyberphysical digital microfluidic biochips. *IEEE Trans. CAD Integr. Circuits Sys.* (1), 32 (2013), 59–72.
- Luo, Y., Chakrabarty, K., Ho, T. Real-time error recovery in cyberphysical digital-microfluidic biochips using a compact dictionary. *IEEE Trans. CAD Integr. Circuits Sys* (12), 32 (2013), 1839–1852.
- Mugele, F., Baret, J. Electrowetting: From basics to applications. *J. Phys.: Condens. Matter*, 17 (2005), 705–R774.
- Mullis, K.B., Erlich, H.A., Arnhem, N.,

- Horn, G.T., Saiki, R.K., Scharf, S.J. *Process for amplifying, detecting, and/or-cloning nucleic acid sequences*. US Patent 4,683,195; July 28 1987.
- Ott, J., Loveless, T., Curtis, C., Lesani, M., Brisk, P. BioScript: Programming safe chemistry on laboratories-on-a-chip. In *Proceedings of OOPSLA '18* (Boston, MA, USA, Nov. 7–9, 2018), Article 124.
- Pollack, M.G., Shenderov, A.D., Fair, R.B. Electrowetting-based actuation of droplets for integrated microfluidics. *Lab on a Chip* (2), 2 (2002), 96–101.
- Synthace. Antha-lang, coding biology. 2016. <https://www.antha-lang.org>. [Accessed: 2016-11-01].
- Urbanski, J.P., Thies, W., Rhodes, C., Amarasinghe, S., Thorsen, T. Digital microfluidics using soft lithography. *Lab Chip*, 6 (2006), 96–104.
- Zhao, Y., Xu, T., Chakrabarty, K. Integrated control-path design and error recovery in the synthesis of digital microfluidic lab-on-chip. *JETC* (3), 6 (2010), 11:1–11:28.

Jason Ott, Tyson Loveless, Chris Curtis, Mohsen Lesani, and Philip Brisk ([jott002, tlove004, ccurt002]@ucr.edu,

[lesani, philip]@cs.ucr.edu), University of California, Riverside, CA, USA.

© 2021 ACM 0001-0782/21/2 \$15.00

# Digital Threats: Research and Practice (DTRAP)

Open for Submissions

*A peer-reviewed journal that targets the prevention, identification, mitigation, and elimination of digital threats*



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

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



# Technical Perspective

## Solving the Signal Reconstruction Problem at Scale

By Zachary G. Ives

WHEN PROBLEMS ARE scaled to “big data,” researchers must often come up with new solutions, leveraging ideas from multiple research areas—as we frequently witness in today’s big data techniques and tools for machine learning, bioinformatics, and data visualization. Beyond these heavily studied topics, there exist other classes of general problems that must be rethought at scale. One such problem is that of *large-scale signal reconstruction*:<sup>4</sup> taking a set of observations of relatively low dimensionality, and using them to reconstruct a high-dimensional, unknown signal. This class of problems arises when we can only observe a subset of a complex environment that we are seeking to model—for instance, placing a few sensors and using their readings to reconstruct an environment’s temperature, or monitoring

**The following paper is notable because it scalably addresses an underserved problem with practical impact, and does so in a clean, insightful, and systematic way.**


multiple points in a network and using the readings to estimate end-to-end network traffic, or using 2D slices to reconstruct a 3D image.

This *signal reconstruction problem* (SRP) is typically approached as an optimization task, in which we search for the high-dimensional signal that minimizes a loss function comparing it to the known properties of the signal. Prior solutions to the SRP make use of linear algebra techniques<sup>4</sup> or expectation maximization<sup>2</sup> to find a solution. However, at scale, the dimensionality of the signal is high enough to render such optimization techniques too costly. In the following paper, Asudeh et al. show that algorithmic insights about SRP, combined with database techniques such as similarity joins and sketches, can be used to scalably solve the signal reconstruction problem. The paper creatively integrates query processing, approximation, and linear algebra techniques.

The authors start by noting that SRP is a special case of quadratic programming, which they exploit by solving the Lagrangian dual formulation of the original problem. Building upon this, they make a connection to query processing: the key part of the algorithm computes the product of a (typically very sparse) matrix  $A$  with its transpose,  $AA^T$ . In turn, that computation derives most of its value from a small number of elements from  $A$ .

The authors creatively leverage this observation to handle huge matrices by implementing matrix multiplication via a set-intersection primitive. They build upon set-similarity joins and apply threshold-based techniques<sup>3</sup> to bound the values of the matrix product, thus developing a fast approximation algorithm. Finally, they show how to use min-hash

sketches<sup>1</sup> to approximate the sets, allowing further trade-offs of accuracy vs performance (and space). Experimental analysis shows these techniques scale well enough to predict end-to-end routes in a large P2P network, which is several orders of magnitude larger than prior solutions could handle.

This paper is notable because it scalably addresses an underserved problem with practical impact, and does so in a clean, insightful, and systematic way. It makes several key contributions. First, it shows how insights into the linear algebra computation can be used for greater efficiency (the connection to quadratic programming, which allows it to be solved via the Lagrangian dual). Subsequently, it makes insightful connections to techniques from query processing and sketches to develop approximation algorithms. Finally, the authors conduct an experimental study demonstrating high performance at scale. They illustrate the potential benefits of connecting important optimization problems with database approximate query processing techniques. 

### References

1. Broder, A. On the resemblance and containment of documents. *Compression and Complexity of Sequences*. IEEE (1997), 21–29.
2. Cao, J., Davis, D., Wiel, S.V. and Yu, B. Time-varying network tomography: Router link data. *J. American Statistical Assoc.* 95, 452 (2000), 1063–1075.
3. Chaudhuri, S., Ganti, V. and Kaushik, R. A primitive operator for similarity joins in data cleaning. *ICDE*. IEEE (2006), 5.
4. Vogel, C.R. Computational methods for inverse problems *SIAM* 23 (2002).

**Zachary G. Ives** is Department Chair and Adani President’s Distinguished Professor of Computer and Information Science at the University of Pennsylvania, Philadelphia, PA, USA.

Copyright held by author.

# Scalable Signal Reconstruction for a Broad Range of Applications

By Abolfazl Asudeh, Jeess Augustine, Saravanan Thirumuruganathan, Azade Nazi, Nan Zhang, Gautam Das, and Divesh Srivastava

## Abstract

**Signal reconstruction problem (SRP) is an important optimization problem where the objective is to identify a solution to an underdetermined system of linear equations that is closest to a given prior. It has a substantial number of applications in diverse areas, such as network traffic engineering, medical image reconstruction, acoustics, astronomy, and many more. Unfortunately, most of the common approaches for solving SRP do not scale to large problem sizes. We propose a novel and scalable algorithm for solving this critical problem. Specifically, we make four major contributions. First, we propose a dual formulation of the problem and develop the DIRECT algorithm that is significantly more efficient than the state of the art. Second, we show how adapting database techniques developed for scalable similarity joins provides a substantial speedup over DIRECT. Third, we describe several practical techniques that allow our algorithm to scale—on a single machine—to settings that are orders of magnitude larger than previously studied. Finally, we use the database techniques of materialization and reuse to extend our result to dynamic settings where the input to the SRP changes. Extensive experiments on real-world and synthetic data confirm the efficiency, effectiveness, and scalability of our proposal.**

## 1. INTRODUCTION

The database community has been at the forefront of grappling with challenges of big data and has developed numerous techniques for the scalable processing and analysis of massive datasets. These techniques often originate from solving core data management challenges but then find their way into effectively addressing the needs of big data analytics. We study how database techniques can benefit *large-scale signal reconstruction*,<sup>13</sup> which is of interest to research communities as diverse as computer networks,<sup>15</sup> medical imaging,<sup>7</sup> etc. We demonstrate that the scalability of existing solutions can be significantly improved using ideas originally developed for similarity joins<sup>5</sup> and selectivity estimation for set similarity queries.<sup>3</sup>

**Signal reconstruction problem (SRP):** The essence of SRP is to solve a linear system of the form  $AX = b$ , where  $X$  is a high-dimensional unknown *signal* (represented by an  $m$ -d vector in  $\mathbb{R}^m$ ),  $b$  is a low-dimensional projection of  $X$  that can be observed in practice (represented by an  $n$ -d vector in  $\mathbb{R}^n$  with  $n \ll m$ ), and  $A$  is an  $n \times m$  matrix that captures the linear relationship between  $X$  and  $b$ . There are many real-world applications that follow the SRP model (see Section 2.1). High-dimensional signals such as environmental temperature can only be observed through low-dimensional

observations, such as readings captured by a small number of temperature sensors. End-to-end network traffic, another high-dimensional signal, is often monitored through low-dimensional readings such as traffic volume on routers in the backbone or edge networks. In these applications, the laws of physics or the topology of computer networks reveal the value of  $A$ , and our objective is to reconstruct the high-dimensional signal  $X$  from the observation  $b$  based on the knowledge of  $A$ .

As  $n \ll m$ , the linear system is underdetermined. That is, for a given  $A$  and  $b$ , there are an infinite number of feasible solutions (of  $X$ ) that satisfy  $AX = b$ . In order to identify the best reconstruction of the signal, it is customary to define and optimize for a *loss function* that measures the distance between the reconstructed  $X$  and a prior understanding of certain properties of  $X$ . For instance, one's prior belief of  $X$  can be specified as an  $m$ -d vector  $X'$  and define the loss function as the  $\ell_2$ -norm of  $X - X'$ , that is,  $\|X - X'\|_2$ . In other cases, when prior knowledge indicates that  $X$  is sparse, one can define the loss function as the  $\ell_2$ -norm of  $X$ , aiming to minimize the number of nonzero elements in the reconstructed signal. For the purpose of this paper, we consider the  $\ell_2$ -based loss function of  $\|X - X'\|_2$ , which has been adopted in many application-oriented studies such as Grangeat and Amans<sup>7</sup> and Zhang et al.<sup>15</sup>

**Running example of SRP:** SRP has a broad range of applications. For the ease of exposition, we use as a running example based on network tomography (Section 2.1), where the objective is to compute the pairwise end-to-end traffic in IP networks. Pairwise traffic measures the volume of traffic between all pairs of source-destination nodes in an IP network and has numerous uses such as capacity planning, traffic engineering, and detecting traffic anomalies. Informally, consider an IP network where various sources and destinations send different amounts of traffic to each other. The network administrator is aware of the network topology and the routing table (from which we can construct matrix  $A$ ). In addition, the administrator can observe the traffic passing through each link in the backbone network (observation  $b$ ). The goal is to find the amount of traffic flow between all source-destination pairs (signal  $X$ ). Note that one cannot directly measure the raw traffic between all source-destination pairs due to challenges in

The original version of this paper was entitled "Leveraging Similarity Joins for Signal Reconstruction" and was published in *PVLDB 10*, 11 (2018), 1276–1288.



instrumentation and storage—see Zhang et al.<sup>15</sup> for a technical discussion. In almost all real-world IP networks, the number of source-destination pairs is significantly larger than the number of links, leading to an underdetermined linear system. To reconstruct the pairwise traffic, the network community introduced various traffic models, for example, the gravity model,<sup>15</sup> as the prior for  $X'$ , and used the  $\ell_2$ -distance between  $X$  and the prior as the loss function. Note that in reconstructing the pairwise distances, efficiency is a concern front and center, especially given the rise of software designed networks (SDNs) that feature much larger sizes and much more frequent topological changes, pushing further the scalability requirements of signal reconstruction algorithms.

**Research gap:** Because of the importance of SRP, there has been extensive work from multiple communities on finding efficient solutions. To solve the problem efficiently, methods explored in the recent literature include statistical likelihood-based iterative algorithms based on expectation-maximization, as well as the use of linear algebraic techniques such as computing the pseudoinverse of  $A$ <sup>13</sup> or performing singular value decomposition (SVD) on  $A$ , and iterative algorithms for solving the linear system.<sup>13</sup> Yet even these approaches cannot scale to fully meet the requirements in practice, especially in settings such as traffic reconstruction in large-scale IP networks—which call for a more scalable solution.

**Our approach:** In this paper, we consider a special case of SRP where  $A$ ,  $X$ , and  $b$  are nonnegative with  $A$  being a *sparse binary matrix*. Such a setting finds its applications in many domains, as explained in Section 2.1. We present an exact algorithm (DIRECT) based on the transformation of the problem into its Lagrangian dual representation. DIRECT already outperforms commonly used approaches for SRP, as it avoids expensive linear algebraic operations required by the previous solutions and scales up to medium-size settings. Next, we investigate whether our approach can be sped up even further, by replacing exact computations with approximation techniques. After a careful investigation of DIRECT, it turns out that the computational bottleneck is a special case of matrix multiplication involving a sparse binary matrix with its transpose. We use the observation that a small number of cells in the result matrix of the bottleneck operation take the bulk of the values and propose a threshold-based algorithm for approximating it. Specifically, we reduce the problem to computing the dot product of two vectors if and only if their similarity is above a user-provided threshold. Our key idea here is to leverage various database techniques to speed up the multiplication operation. We propose a hybrid algorithm based on a number of techniques originally proposed for computing similarity joins and selectivity estimation of set similarity queries, resulting in significant speedup, enabling our proposal to scale to large-scale settings.

We push the boundaries to very large systems (VLS) with sizes in the order of a million equations with a billion unknowns. We identify that the barrier to this extension is the output size of the multiplication of  $A$  by its transpose,

that is,  $AA^T$ , as it is simply too large to be kept in memory. We conducted careful theoretical analyses and experimental evaluation on the number of nonzero elements in this matrix that confirm the matrix is sparse in practice. We then leverage this sparsity to efficiently solve very large systems of equations. Finally, we consider the scenario where the input to our problem changes dynamically. We pay attention to the observation that the underlying structure of the system  $A$  does not change frequently. Vector  $b$ , on the other hand, may change often. We utilize the database technique of materialization and reuse a carefully constructed *signature matrix* for dynamic settings.

## 2. PROBLEM FORMULATION

We consider a special class of SRP that has a number of applications in network traffic engineering, tomographic image reconstruction, and many others. We are given a system of linear equations  $AX = b$  where

- $A \in \{1, 0\}^{n \times m}$  is a sparse binary matrix  $n \ll m$ .
- $X \in \mathbb{R}^m$  is the “signal” to be reconstructed and is a vector of unknown values.
- $b \in \mathbb{R}^n$  is the vector of observations.

Each row in the matrix  $A$  corresponds to an equation with each column corresponding to an unknown variable. When the number of equations ( $n$ ) is much smaller than the number of unknowns ( $m$ ), the system of linear equations is said to be underdetermined and does not have a unique solution. The solution space can be represented as a hyperplane in an  $m' \in [2, m]$  dimensional vector space.<sup>a</sup> Because SRP does not have a unique solution, one must have auxiliary criteria to choose the best solution from the set of (possibly infinite) valid solutions. A common approach in SRP is to provide a prior  $X'$  and the objective is to pick the solution  $X$  that is closest to  $X'$ . We study the problem where the objective is to find the point satisfying  $AX = b$  that minimizes the  $\ell_2$ -distance from a prior point  $X'$ . Formally, the problem is defined as:

$$\begin{aligned} \min \|X - X'\|_2 \\ \text{s.t. } AX = b \end{aligned} \quad (1)$$

<sup>a</sup> We assume that the problem has at least one solution.

**Figure 1. Visualizing the problem.**

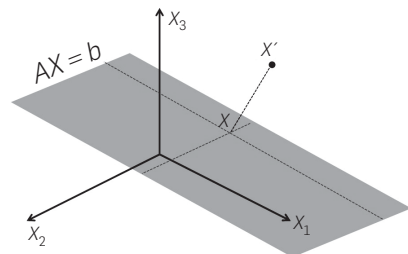


Figure 1 provides a visualization of the problem in three dimensions. The gray plane is the solution space with the prior marked as a point  $X'$ . The intersection of the perpendicular line to the plane that passes through  $X'$  is the point that minimizes  $\|X - X'\|_2$ .

We observe that SRP is a special case of quadratic programming where (a) the constraints are only in the form of equality, (b) matrix  $A$  is sparse, and (c) matrix  $A$  is binary (and hence unweighted). By leveraging these characteristics, we seek to design more efficient solutions compared with the baselines that are designed for general cases. In Section 3, we use the dual representative of the problem to propose an efficient exact algorithm. In Section 4, we show how leveraging similarity joins techniques help in achieving significant speedup without sacrificing much accuracy.

## 2.1. Applications of SRP

SRP covers a broad range of real-world problems that use signal reconstruction. In practice, it is popular to observe low-dimensional projections in the form of (unweighted) aggregates of a high-dimensional signal vector. For example, in general network flow applications (such as road traffic estimation<sup>16</sup>), the value on each edge is the summation of the flow values that includes this edge as part of the path between them. Of course, a requirement to our problem is an “expert-provided” prior template, such as *gravity model*<sup>15</sup> for the network flow problems. Another major application domain for SRP problem over aggregates is image reconstruction, where observations are unweighted projections of unknowns. Image reconstruction has broad applications ranging from medical imaging<sup>7</sup> to astronomy<sup>14</sup> and physics.<sup>10</sup> Some of the other applications of SRP, in general, include radar data reconstruction<sup>9</sup> and transmission electron microscopy,<sup>8</sup> to name a few. To showcase some applications in more detail, we sketch a few examples in the context of network flow problems and image reconstruction in the following.

**Network tomography. Traffic matrix computation (the running example):** Consider an IP network with  $n$  traffic links and  $m$  source-destination traffic flows (SD flow) between the ingress and egress points, where  $n \ll m$ . The ingress/egress points can be points of presence (PoPs) or routers or even IP prefixes depending on the level of granularity required. The network has a routing policy and prescribes a path for each of the SD flows that can be captured in a  $\#links(n) \times \#flows(m)$  binary matrix  $A$ , where the entry  $A[i, j] = 1$  if the link  $i$  is used to route the traffic of the  $j$ th SD flow. The matrix  $A$  is sparse and “fat” with more SD flows (columns) than number of links (rows). Note that, one cannot directly measure each of the SD flows on a link owing to efficiency reasons. However, one can easily measure the total volume of the network traffic that passes through a given link using network protocols such as SNMP. Thus, the load on each link  $i$  becomes the observed vector  $b$ . To obtain a prior  $X'$ , one can use any traffic model such as the popular and intuitive *gravity model*.<sup>15</sup> It assumes independence between source and destination and states that traffic between any given source  $s$  and destination  $d$  is proportional to the product of network traffic entering at  $s$  and that exiting at  $d$ .

**Traffic analysis attack in P2P networks:** In traffic analysis attack, the information leak on traffic data is exploited to expose the user traffic pattern in P2P networks. Here, we propose the following traffic analysis attack that can be modeled to our problem: consider an adversary who monitors the link level traffics in a P2P network. Applying SRP, one can directly identify the volume of traffic between any pair of users in a P2P network.

**Image reconstruction.** Image reconstruction<sup>7</sup> has a wide range of applications in different fields such as medical imaging,<sup>7</sup> and physics.<sup>10</sup> Given a set of (usually 2D) projection of a (usually 3D) image, the objective is to reconstruct it. The reconstruction is usually done with the help of some prior knowledge. For example, knowing that the 2D projections are taken from a human face, one may use a template 3D face photo and, among all possible 3D reconstructions from the 2D images, find the one that is the closest to the template, making the image reconstruction more effective.

**CT scan:** A popular application of SRP is tomographic reconstruction, which is a multidimensional linear inverse problem with wide range of applications in medical imaging<sup>7</sup> such as CT scans (computed tomography). A CT scan takes multiple 2D projections (vector  $b$ ) through X-rays from different angles (matrix  $A$ ) and the objective is to reconstruct the 3D image from the projections. Many 3D images may produce the same projections necessitating the use of priors to choose an appropriate reconstruction.

**Radio astronomy:** In astronomy, SRP has application for reconstructing interferometric images where the astrophysical signals are probed through Fourier measurements. The objective is to reconstruct the images from the observations—forming an SRP scenario. Also, the specific prior information about the signals plays an important role in reconstruction, as mentioned in Wiaux et al.<sup>14</sup>

## 3. EXACT SOLUTION FOR SOLVING SRP

We begin by describing two representative approaches for solving SRP from prior research and highlight their shortcomings. We then propose a dual representation of the problem that can be solved exactly in an efficient manner and already outperforms the baselines. This alternate formulation allows one to leverage various database techniques for speeding it up.

### 3.1. Lagrangian formulation of SRP

We leverage the Lagrangian dual form of SRP as a special case of quadratic programming and design an efficient exact solution for it. For SRP as specified in Equation 1,  $f(X) = \frac{1}{2} X^T X - X'^T X$  and  $g(X) = AX$ .<sup>b</sup> Thus, our problem can be rewritten as:

$$L(X, \lambda) = \frac{1}{2} X^T X - X'^T X + \lambda^T (AX - b) \quad (2)$$

<sup>b</sup> Note that  $\min \frac{1}{2} X^T X - X'^T X$  is the same as  $\min \|X - X'\|_2$ .

<sup>c</sup> Because, looking at Figure 1, Equation 1 has a single optimal point, Equation 2 has one stationary point that happens to be the saddle point.

Next, we find the stationary point<sup>c</sup> of Equation 2 in the general form by taking the derivatives with regard to  $X$  and  $\lambda$ , and setting them to zero, we get:

$$X = X' - A^T(AA^T)^{-1}(AX' - b) \quad (3)$$

**Solving SRP in dual form.** The stationary point of Equation 2 is the optimal solution for our problem (Equation 1). In contrast to prior work, we solve the SRP problem by directly solving Equation 3. We make two observations. First, the matrix  $AA^T \in \mathbb{Z}^{n \times n}$  always has an inverse as it is full rank. From Figure 1, one can note that the problem has a unique solution that minimizes the distance from the prior. It means that  $AA^T$  is full rank, because otherwise the problem was not feasible and would not have a solution. Second, Equation 3 does have a matrix inverse operator that is expensive to compute. However, one can avoid taking the inverse of  $AA^T$  by computing  $\xi$  in Equation 4 and replacing  $(AA^T)^{-1}(AX' - b)$  by it in Equation 3.

$$(AA^T)\xi = AX' - b \quad (4)$$

Algorithm 1 provides the pseudocode for DIRECT.

**Algorithm 1** DIRECT

**Input:**  $A$ ,  $b$ , and  $X'$

**Output:**  $X$

- 1:  $t = AA^T$
- 2:  $t_2 = AX' - b$
- 3: Solve system of linear equations:  $t \xi = t_2$
- 4:  $X = X' - A^T \xi$
- 5: **return**  $X$

**Performance analysis of DIRECT.** Let us now investigate the performance of our algorithm. Recall that  $A$  is a fat matrix with  $n \ll m$ , whereas  $X$  and  $X'$  are  $m$ -dimensional vectors, and  $b$  is a  $n$ -dimensional vector. Line 1 of Algorithm 1 takes  $O(n^2m)$ , whereas Line 2 takes  $O(nm)$ . Line 3 involves solving a system of linear equations. A naive way would be to compute the inverse of  $t$  that can take as much as  $O(n^3)$ . However, by observing that  $t$  is sparse, one can use approaches such as Gauss-Jordan elimination or other iterative methods that are practically much faster for sparse matrices. Finally, the computation of Line 4 is in  $O(nm)$ . Looking at DIRECT holistically, one can notice that its computational bottleneck is Line 1, thereby making the overall complexity to be  $O(n^2m)$ .

An additional approach to speedup DIRECT is to observe that matrix  $A$  is sparse and thereby to store

**Figure 2. Illustration of the sparse representation of  $A$ . (a) Nonsparse representation and (b) sparse representation.**

0	0	0	1	0	0	0	1	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	1	0	1	0	1
0	1	0	0	0	0	1	0	0	0

(3, 7)
(2)
(5, 7, 9)
(1, 6)

(a)
(b)

it in a manner that allows efficient matrix multiplication. Because  $A$  is binary (and hence unweighted), a natural representation is to store only the indices of nonzero values. Figures 2a and 2b show the nonsparse and sparse representation of a matrix  $A$ . Note that  $AA^T$  is symmetric as  $t[i, j]$  and  $t[j, i]$  are obtained by the dot product of rows  $i$  and  $j$  of  $A$ . Let  $l$  be the number of nonzero elements in each row. Because  $A$  is sparse,  $l \ll m$ , one can design a natural matrix multiplication algorithm with time complexity of  $O(nml)$  that is orders of magnitude faster than algorithm such as Strassen algorithm.

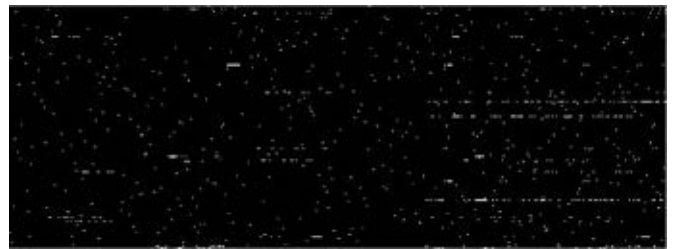
**4. TRADING OFF ACCURACY WITH EFFICIENCY**

In many applications of SRP,  $m$  is often in  $O(n^2)$ , thereby making the computational complexity of DIRECT to be  $O(n^4)$ . The key bottleneck is the computation of  $AA^T$ . On the other hand, for large problem instances, the user may accept trading off accuracy with efficiency and prefer a close-to-exact solution that is computed quickly, rather than the expensive exact solution. Our objective is to speed up DIRECT by computing the bottleneck step, that is, computing  $AA^T$ , approximately. We show how to leverage a threshold-based approach by only computing the values of matrix  $AA^T$  that are larger than a certain threshold. We describe the connection between this problem variant and similarity joins and propose a hybrid method by adopting two classical algorithms designed for similarity estimation, which results in an efficient solution for computing  $AA^T$ .

**4.1. Bounding values in matrix  $AA^T$**

We begin by showing that one can efficiently compute the bound for each cell value in matrix  $AA^T$ . Figure 3 shows a sparse matrix  $A$  with 183 rows and 495 columns, in which the

**Figure 3. An example of the binary sparse matrix  $A_{183 \times 495}$ .**



**Figure 4. The nonzero elements in  $AA^T$  for the example of Figure 3.**





nonzero elements are highlighted in white. Figure 4 shows the nonzero elements in matrix  $AA^T$ . We can notice that  $AA^T$  is square and also sparse due to the fact that every element of  $AA^T$  is the dot product of two sparse vectors (two rows of matrix  $A$ ). Furthermore, one can also observe a more subtle phenomenon that we state in Theorem 1, which could be used to design an efficient algorithm.

**THEOREM 1.** *Given a sparse binary matrix  $A$ , considering the elements on the diagonal of  $AA^T$ , that is,  $t[i, i], \forall 0 \leq i < n$ :*

- $t[i, i] = |A[i]|$ , where  $|A[i]|$  is the number of nonzero elements in row  $A[i]$ .
- $t[i, i]$  is an upper bound for the elements in the row  $t[i]$  and the column  $t[, i]$ ; formally,  $\forall 0 \leq j < n: t[i, j] \leq t[i, i]$  and  $t[i, j] \leq t[j, j]$ .

The proof can be found in Asudeh et al.<sup>2</sup>

Consider two representations of  $AA^T$  of the example matrix given in Figure 3. Figure 4 shows all the nonzero elements of  $AA^T$ , whereas Figure 5 shows a magnitude-weighted variant wherein cells with larger values are plotted in brighter colors. Figure 5 visually shows that the elements on the diagonal are brighter than the ones in the same row and column as predicted by Theorem 1. One may notice that most of the nonzero elements of  $AA^T$  (in Figure 4) are small values (in Figure 5). Although there are a reasonable number of nonzero elements, the number of elements with higher magnitude is often much smaller. Next, we use this insight along with Theorem 1 for speeding up DIRECT.

#### 4.2. Threshold-based computation of $AA^T$

By developing a bound on the cell values in  $AA^T$ , we can see that a small number of elements in  $AA^T$  take the bulk of the value. This is the key in designing a threshold-based algorithm for computing  $AA^T$  wherein we only compute values of  $AA^T$  that are above a certain threshold. Specifically, we use the elements on the diagonal as an upper bound and only compute the elements for which this upper bound is larger than a user-specified threshold. Note that, if the threshold is equal to 1, the algorithm will compute the values of all elements. However, the user-specified threshold allows additional opportunities for efficiency.

Algorithm 2 provides the pseudocode for the threshold-based multiplication of sparse binary matrix  $A$  with its transpose. This algorithm depends on the existence of an oracle called SIM that given two rows  $A[i]$  and  $A[j]$ , and the threshold  $\tau$ , returns the dot product of  $A[i]$  and  $A[j]$  if the result is not less than  $\tau$ .

---

#### Algorithm 2 Approx $AA^T$

**Input:** Sparse matrix  $A$ , Threshold  $\tau$

**Output:**  $t$

---

- 1:  $\mathcal{F} = \{\}$
- 2: **for**  $i = 0$  to  $n - 1$  **do**
- 3:  $t[i, i] = |A[i]|$
- 4: **if**  $|A[i]| \geq \tau$  **then** add  $i$  to  $\mathcal{F}$

**Figure 5. Magnitude of weights in  $AA^T$  for the example of Figure 3.**



- 5: **end for**
  - 6: **for** every pair  $i, j \in \mathcal{F}$  **do**
  - 7:  $t[i, j] = t[j, i] = \text{SIM}(A[i], A[j], \tau)$
  - 8: **end for**
  - 9: **return**  $t$
- 

#### 4.3. Leveraging similarity joins for Oracle SIM

The database community has extensively studied mechanisms for computing set similarity for applications such as data cleaning<sup>5</sup> where the objective is to efficiently identify the set of tuples that are “close enough” on multiple attributes. We next describe how to implement the oracle SIM by leveraging prior research on computing set similarity. Especially, we propose a hybrid method that combines the threshold-based similarity joins with the sketch-based methods to resolve their shortcomings.

**Oracle SIM through set similarity.** Given two rows  $A[i]$  and  $A[j]$ , and the threshold  $\tau$ , SIM should find the dot product of  $A[i]$  and  $A[j]$  if it is not less than  $\tau$ . We can make an interesting connection between SIM and set similarity problems as follows. Let every column in matrix  $A$  be an object  $o$  in a universe  $\mathcal{U}$  of  $m$  elements. Every row  $A[i]$  represents a set  $U_i$  in  $\mathcal{U}$ , where  $\forall o_j \in \mathcal{U}, o_j \in U_i$  iff  $A[i, j] = 1$ . Equivalently, each row corresponds to a set  $U_i$  that stores the indices of the nonzero columns similar to Figure 2b. Using this transformation, we can see that our objective is to compute  $|U_i \cap U_j|$  for all pairs of sets  $U_i$  and  $U_j$  where  $|U_i \cap U_j| \geq \tau$ . Note that we represent  $|U_i \cap U_j|$  by  $\cap_{ij}$  and  $|U_i \cup U_j|$  by  $\cup_{ij}$ , respectively.

Due to its widespread importance, different versions of this problem have been extensively studied in the DB community. We consider one exact approach and two approximate approaches based on threshold-based algorithms<sup>5</sup> and sketch-based methods.<sup>3,6</sup> We then compare and contrast the two approximate approaches, describe the scenarios when they provide better performance, and propose a hybrid algorithm based on these scenarios.

**Exact approach: set intersection.** One can see that when  $\tau = 1$ , the problem boils down to computing  $AA^T$  exactly. This in turn boils down to computing the intersection between two sets as efficiently as possible. The sparse representation of the matrix often provides the nonzero columns in an ordered manner. The simplest approaches for finding the intersection of ordered sets is to perform a linear merge by scanning both the lists in parallel and leveraging the ordered nature similar to the merge step of merge sort. One can also speedup this approach by using sophisticated approaches

such as binary search on one of the lists or using sophisticated data structures such as treaps or skip lists. Each of these approaches allows one to “skip” some elements of a set when necessary.

**Approximate approach: threshold-based algorithms.** Threshold-based algorithms, such as Chaudhuri et al.,<sup>5</sup> identify the pair of sets such that their similarity is more than a given threshold. This has a number of applications such as data cleaning, deduplication, collaborative filtering, and product recommendation in advertisement where the objective is to quickly identify the pairs that are highly similar. The key idea is that if the intersection of two sets is large, the intersection of small subsets of them is nonzero.<sup>5</sup> More precisely, for two sets  $U_i$  and  $U_j$  with size  $h$ , if  $\cap_{i,j} \geq \tau$ , any subset  $U'_i \subset U_i$  and  $U'_j \subset U_j$  of size  $h - \tau + 1$  will overlap; that is,  $|U'_i \cap U'_j| > 0$ . Using this idea, while considering an ordering of the objects, the algorithm first finds the set of candidate pairs that overlap in a subset of size  $h - \tau + 1$ . In the second step, the algorithm verifies the pairs, by removing the false positives.

One can see that the effectiveness of this method highly depends on the value of  $\tau$  and, considering the target application, it works well for the cases where  $\tau$  is large. For example, consider a case where  $h = 100$ . When  $\tau = 99$  (i.e., 99% similarity), the first filtering step needs to compare the subsets of size 2 and is efficient, whereas if  $\tau = 10$ , the filtering step needs to compare the subset pairs of size 91, which is close to the entire set. The latter case is quite possible in our problem. To understand it better, let us consider matrix  $A$  in Figure 3, while setting  $\tau$  equal to 5 in Algorithm 2. Even though the size of many of the rows is close to the threshold, there are rows  $A[i]$  where  $|A[i]|$  is significantly larger than it. For example, for two rows  $A[i]$  and  $A[j]$  where  $|A[i]| \geq 50$  and  $|A[j]| \geq 50$ , to satisfy the condition that the dot product should not be less than  $\tau$ , the filtering step needs to compare the subsets of size  $\geq 44$ , which is close to the exact comparison of  $A[i]$  and  $A[j]$ .

**Approximate approach: sketch-based algorithms.** Sketch-based methods such as Beyer et al.<sup>3</sup> and Cohen and Kaplan<sup>6</sup> use a precomputed synopsis such as a minhash for answering different set aggregates such as Jaccard similarity. The main idea behind the minhashing-based algorithms is as follows: consider a hash (ordering) of the elements in  $\mathcal{U}$ . For each set  $U_i$ , let  $h_{\min}(U_i)$  be the element  $o \in U_i$  that has the minimum hash value. Two sets  $U_i$  and  $U_j$  have the same minhash, when the element with the smallest hash value belongs to their intersection. Hence, it is easy to see that the probability that  $h_{\min}(U_i) = h_{\min}(U_j)$  is equal to  $\frac{\cap_{i,j}}{h}$ , that is, Jaccard similarity of  $U_i$  and  $U_j$ . Bottom- $k$  sketch,<sup>6</sup> a variant of minhashing, picks the hash of the  $k$  elements in  $U_i$  with the smallest hash value, as its signature. The Jaccard similarity of two sets  $U_i$  and  $U_j$  is estimated as  $\frac{k_{\cap}(i,j)}{k}$ , where  $k_{\cap}(i,j)$  is  $|h_k(U_i) \cap h_k(U_j)|$ . Beyer et al.<sup>3</sup> use the bottom- $k$  sketch for estimating the union and intersection of the sets. Let  $h_{i,j}[k]$  be the hash value of the  $k$ th smallest hash value in  $h_k(U_i) \cup h_k(U_j)$ . The idea is that the larger the size of a set is, the smaller the expected

value of the  $k$ th element in hash is. Using the results of Beyer et al.,<sup>3</sup>  $\frac{k_{\cap}(i,j)}{h_{i,j}[k]}$  is an unbiased estimator for  $\cap_{i,j}$ . Hence, the estimation for  $\cap_{i,j}$  is as provided in Equation 5.

$$E[\cap_{i,j}] = \frac{k_{\cap}(i,j)}{k} \frac{m(k-1)}{h_{i,j}[k]} \quad (5)$$

Estimating  $\cap_{i,j}$  with Equation 5 performs well when  $\cap_{i,j} \gg 1$ ,<sup>3</sup> that is, the larger sets. Hence, we combine the threshold-based and sketch-based algorithms to design the oracle SIM, as a hybrid method that, based on the sizes of the rows  $A[i]$  and  $A[j]$ , adopts the threshold-based computation with sketch-based estimation for computing the dot product of  $A[i]$  and  $A[j]$ . We consider  $\log(m)$  as the threshold to decide which strategy to adopt. Considering the effectiveness of threshold-based approaches when  $U_i$  and  $U_j$  are small and, as a result, the two sets need a large overlap to have the intersection larger than  $\tau$ , if  $|U_i|$  and  $|U_j|$  are less than  $\log(m)$ , we choose the threshold-based intersection computation. However, if the size of  $U_i$  or  $U_j$  is more, then we use the bottom- $k$  sketch, while considering  $k$  to be  $\log(m)$ . For each element  $o_j \in \mathbb{U}$ , we set  $h(o_j) = j$ . Hence, for each vector  $U_i$ , the index of the first  $\log(m)$  elements in it is its bottom- $k$  sketch. Using this strategy, Algorithm 3 shows the pseudocode of the oracle SIM.

Given two sets  $U_i$  and  $U_j$  (corresponding to the rows  $A[i]$  and  $A[j]$ ) together with the threshold  $\tau$ , the algorithm aims to compute the value of  $\cap_{i,j}$ , if it is larger than  $\tau$ . Combining the two aforementioned methods, if  $|U_i|$  and  $|U_j|$  are more than a value  $\alpha$ , the algorithm uses sampling to estimate  $\cap_{i,j}$ ; otherwise, it applies the threshold-based method to compute it. During the sampling, rather than sampling from  $\mathcal{U}$ , the algorithm samples from  $U_i$  to reduce the underestimation of probability. In this case, in order to compute  $\cap_{i,j}$ , the algorithm, for each sample, picks a random object from  $U_i$  and checks its existence in  $U_j$ . It is easy to see it is an unbiased estimator for  $\cap_{i,j}$ , where its expected value is  $\cap_{i,j}$ . If  $|U_i|$  or  $|U_j|$  is less than  $\alpha$ , the algorithm applies threshold-based strategy for computing  $\cap_{i,j}$ . As discussed earlier in this subsection, in order for  $\cap_{i,j}$  to be more than  $\tau$ , the subsets of size  $\cap_{i,j} - \tau + 1$  should intersect. Hence, the algorithm first applies the threshold filtering, and only if the two subsets intersect, it continues with computing  $\cap_{i,j}$ .

---

#### Algorithm 3 SIM

**Input:** The sets  $U_i$  and  $U_j$ , Threshold  $\tau$

**Output:**  $c$

---

- 1: **if**  $|U_i| \geq \log(m)$  and  $|U_j| \geq \log(m)$  **then**
- 2:    $h_i$  = the first  $k$  elements in  $U_i$
- 3:    $h_j$  = the first  $k$  elements in  $U_j$
- 4:    $k_{\cap}(i,j) = |h_i \cap h_j|$
- 5:    $h_{i,j}[k]$  = the first  $k$  elements in  $h_i \cup h_j$
- 6:    $c = \frac{k_{\cap}(i,j)}{k} \frac{m(k-1)}{h_{i,j}[k]}$
- 7: **else**
- 8:    $c = 0$
- 9:   **if**  $|U_i| > |U_j|$  **then** swap  $U_i$  and  $U_j$

```

10:  $\beta = |U_i| - \tau$ 
11: for  $k = 0$  to  $\beta$  do: if  $U_i[k] \in U_j$  then  $c = c + 1$ 
12: if  $c = 0$  then return 0
13: for  $k = \beta$  to  $|U_i| - 1$  do: if  $U_i[k] \in U_j$  then  $c = c + 1$ 
14: end if
15: return  $c$ 

```

## 5. SCALING TO VERY LARGE SETTINGS

So far, we considered the scenario where  $n$  is not a large number. Recall that  $n$  is the size of the low-dimensional projection of the unknown variables. We relax this assumption and extend DIRECT for handling the cases where  $n$  is very large (and still  $n \ll m$ ). For example,  $n$  can be in the order of a million, whereas  $m$  is in the order of a billion. A key aspect of DIRECT is that it leverages the sparse representation of the matrix (as against its complete dense representation) for speedup. However, when  $n$  is very large, even fitting the sparse representation of  $A$  into the memory may not be possible. Even if there is only one nonzero value in *every column*, we need  $O(m)$  storage for the matrix.

Interestingly, the similarity joins-based techniques proposed in Section 4 do not require to completely materialize even sparse representation of  $A$  for estimating  $AA^T$ . Also, there are many scenarios where the user is interested in knowing the values of a subset of components of the reconstructed signals such as those corresponding to the largest values of the reconstructed signal. We now show how to adapt our algorithms to handle these scenarios.

Consider Algorithm 1 where the critical step is the first line. Algorithm 3 applies bottom-k sketch for the sets whose size is more than  $\log m$ . Thus, choosing the signature size in the bottom-k sketch to be in  $O(\log m)$ , Algorithm 3 needs at most  $O(\log m)$  elements from *each row*. As a result, Line 1 of DIRECT needs a representation of size  $O(n \log m)$  of  $A$ . For instance, in our example of  $n = 10^6$  and  $m = 10^{12}$ , the size of the representative of  $A$  is only in the order of 1 million rows by 40 columns.

A key assumption for scaling our results to very large settings is that  $t = AA^T$  is sparse in practice. In Asudeh et al.,<sup>1</sup> we theoretically study the sparsity of  $t$ . Specifically, we provide a lower bound and an upper bound on how sparse  $t$  can be. Using an adversarial example, we show the existence of cases for which the matrix is not sparse. Still, as we shall illustrate in Section 7,  $AA^T$  is sparse in practice. It even becomes significantly more sparse after applying thresholding. Therefore, we only store the nonzero values of matrix  $t$ , rather than the complete  $n$  by  $n$  matrix. Line 2 of Algorithm 1 is the multiplication of matrix  $A$  with  $X'$  whose dimensions are  $m$  by 1 followed by subtracting the  $n$ -dimensional result vector from the vector  $b$ . For this line, for each row of  $A$ , we use a sample of size  $O(\log m)$  for the nonzero elements of the row, while using the values of  $X'$  as the sampling distribution. The result is a representation of size  $O(n \log m)$  of  $A$ . Also, rather than loading the complete vector  $X'$  to the memory, in an iterative manner, we bring loadable buckets of it to the memory, update the calculation for that bucket, and move to the next one. In Line 4,  $t$  is the nonzero elements of  $AA^T$  and  $t'$  is an  $n$  by 1 vector, and finding the  $n$  by 1 vector  $\xi$  is doable, using methods such as

Gauss-Jordan. Finally, we only limit the calculations to the variables of interest, or even if the computation of all variables is required, in an iterative manner, we move a loadable bucket of them to the memory, compute their values, and move to the next bucket.

## 6. DYNAMIC SIGNAL RECONSTRUCTION

So far, we focused on the static variant of the SRP, where the objective is to find the point in the answer space that minimizes the distance to the prior  $X'$ . We next investigate a practical scenario where the input to SRP changes. The naive solution is to invoke our algorithms from scratch whenever the input changes. We observed that in many instances of SRP, not all the inputs of a signal reconstruction change. Hence, it is possible to *materialize* some results from the previous iterations and *reuse* them to compute the solution for the current iteration.

Consider our running example of SRP where the objective is to compute the end-to-end traffic between the source-destination pairs in an IP network. Although the actual traffic between the pairs may change quickly, the underlying network topology changes infrequently. Hence, the binary matrix  $A$  is also unchanged. The changes in the traffic affect the observation vector  $b$  and possibly the prior point  $X'$ . Reconstructing the signal  $X$  every few minutes based on current observations when there is no change in network topology is an extremely important scenario in traffic engineering.

Recall that computing  $AA^T$  is the performance bottleneck of DIRECT. Interestingly, because the underlying topology of the graph does not change, the computation of  $AA^T$  can be considered as an amortized preprocessing step that can be materialized and reused for dynamic changes. Also remember that Line 3 of Algorithm 1 uses Gaussian elimination for finding  $\xi$  in  $(AA^T)\xi = AX' - b$ . We observed that this is the second performance bottleneck after the computation of  $AA^T$ . We propose a novel approach<sup>1</sup> to speedup this computation by materializing (and maintaining) an  $n \times (n + 1)$  signature matrix  $S$  that enables the computation of  $\xi$  in  $O(n^2)$ , instead of  $O(n^3)$  for the recomputation.<sup>d</sup>

**Constructing the signature matrix.** For ease of explanation, let  $R = AA^T$  and  $t = AX' - b$ . Now the objective is to find  $\xi$  in:

$$R \xi = t$$

Note that in the above equation,  $R$  is fixed as  $AA^T$  does not change. At a high level, in order to generate the signature matrix, we apply Gaussian elimination for a general form of  $t$  and maintain the delayed operation in the signature matrix. Later on, upon the arrival of an update, we use the signature matrix to updates on  $t$  and compute  $\xi$  accordingly. We would like to highlight the similarity of our signature matrix with LU decomposition, where the matrix  $AA^T$  is decomposed into two matrices  $L$  and  $U$ .<sup>12</sup> Compared to our proposal,<sup>1</sup> the update using the  $L$  and  $U$  matrices as

<sup>d</sup> We note that one can materialize the inverse matrix  $(AA^T)^{-1}$  (or  $A^T(AA^T)^{-1}$ ) as the signature. This however would require more storage and would not give computational advantage compared to our proposal.<sup>1</sup>



signature needs solving of two (albeit specialized) systems of linear equations of time complexity  $O(n^2)$  for computing  $\xi$ . We conducted experiment on validating the effectiveness of our methods and the results are described in Section 7.

## 7. EXPERIMENTAL EVALUATION

### 7.1. Experimental setup

**Hardware and platform.** All of our small-scale experiments were performed on a Macintosh machine with a 2.6 GHz CPU and 8GB memory. The algorithms were implemented using Python 2.7 and MATLAB. For very large setting experiments, we used a 4.0 GHz, 64GB server that runs on Ubuntu 18.04 and the code was rewritten in C++ for scalability and efficiency.

**Datasets.** We conducted extensive experiments to demonstrate the efficacy of our algorithms over graphs with diverse values for a number of nodes, edges, and source-destination pairs. Recall that given a communication network, the size of the routing matrix  $A$  is parameterized by the number of edges and number of source-destination pairs—and not by the number of nodes and edges. We used different datasets with different scales for the experiments. We outline a subset of those datasets in Figure 6. Please refer to Asudeh et al.<sup>1</sup> for the complete list. For small- and medium-size datasets ( $N_1$  in Figure 6), used for comparing against the prior work, we use the synthetic datasets. Our large datasets are real datasets that were derived from a p2p dataset from SNAP repository of Stanford University.<sup>c</sup> Each of the derived large datasets is a subgraph of the overall p2p graph and was obtained by Forest Fire model. For very large (VLS) datasets, we used the complete Gnutella dataset, as well as a popular location-based social networking platform, Brightkite.

Once we sample the network and obtain a connected graph, we consider all possible source destination pairs to be the individual flows. For each source-destination pair, we calculated the shortest path between them and used *Pareto* traffic generation model for generating the flow values. The *prior* point for the experiments ( $X'$ ) was obtained as a function of gravity model from Zhang et al.<sup>15</sup>

### 7.2. Experimental results

We report a representative subset of our experiment results here. Please refer to Asudeh et al.<sup>1,2</sup> for the complete results.

<sup>c</sup> [snap.stanford.edu/data/p2p-Gnutella04.html](http://snap.stanford.edu/data/p2p-Gnutella04.html).

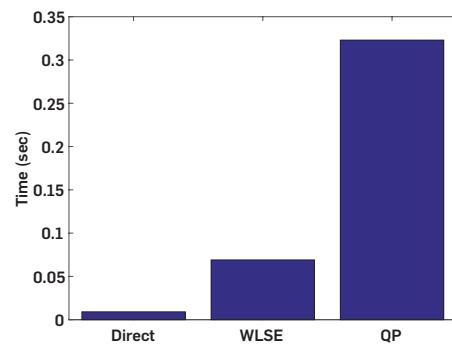
**Figure 6. Dataset characteristics.**

Network	#Nodes	#Edges	#SD pairs
$N_1$	274	281	827
p2p-3	1438	7081	2M
VLS2	10879	44944	32M
VLS3	8298	104469	32M
VLS4	108300	191886	64M
VLS5	36692	372612	128M
VLS6	58228	433106	0.5B

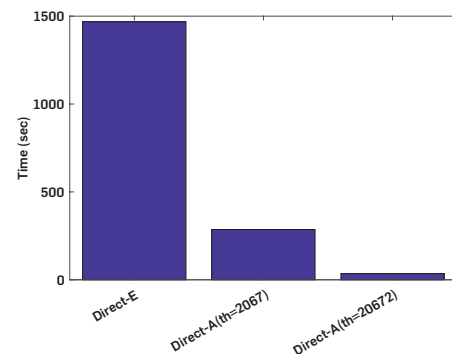
First, as shown in Figure 7, DIRECT significantly outperforms the baselines QP and WLSE<sup>15</sup> on the small dataset  $N_1$ . In addition to comparing with these two baselines, for  $N_1$ , we also used compressive sensing for estimating flow values, which took more than 23 s, even for our smallest setting. We next evaluate the exact version of DIRECT and its approximate counterpart (using Algorithm 2) that leverages techniques from similarity joins to speed up the computation. We use DIRECT-E to refer to the exact version of DIRECT and DIRECT-A for its approximate version. We also evaluate the performance of our algorithms to two different threshold values of  $(m/1000)$  and  $(m/100)$ , where  $m$  is the number of source-destination pairs. Choosing an appropriate threshold is often domain specific with larger thresholds providing better speedups. We compare the performance of the algorithms DIRECT-E and DIRECT-A through execution time and accuracy.

**p2p-3 (2M source-destination pairs).** This network has 2M source-destination pairs with 7081 edges sampled from the *SNAP p2p* dataset. Figure 8 shows that DIRECT-E takes much as 1500 s to compute the exact solution. This is often prohibitive and simply unacceptable for many traffic engineering tasks. However, our approximate algorithms can provide the result in as little as 35 s. This is a significant reduction in execution time with a speedup of as much as 97% over the running time of DIRECT-E. We would like to mention that our experiments<sup>2</sup> demonstrate negligible

**Figure 7. DIRECT v.s. baselines in  $N_1$ :  $n = 281$  and  $m = 827$ .**



**Figure 8. Execution time of DIRECT-E, DIRECT-A( $\tau = 2067$ ), and DIRECT-A( $\tau = 20672$ ) in p2p-3.**



approximation errors, even for threshold value of  $(m/100)$ , which is tolerable for many tasks in network traffic engineering such as routing optimization.<sup>15</sup>

**Sparsity and thresholding results of  $AA^T$ .** We chose VLS2 settings to demonstrate the effectiveness of thresholding, the lower and upper bounds provided by theory for the settings, and an overall reduction in nonzero elements by a suitable threshold. The results are provided in Figure 9. We also included the theoretical lower bound and upper bound in the figure. The number of nonzero values in  $t=AA^T$  for this setting is  $97M$ , which is about 4.85% of the total cells. However, with a modest threshold,  $\tau = 2$ , this number quickly dropped to 0.003%, which highlights the effectiveness of thresholding.

**Scalability results.** In this experiment, we show the scalability of our final algorithm. To do so, we compare the performance of DIRECT-A across different input scales of  $n \times m$ , which confirm the scalability of DIRECT-A for the very large settings through experiments on VLS2 to VLS6. Figure 10 presents the results from scalability experiments for very large settings with varied values of  $n$  and  $m$ . Note that all the experiments are run on a single machine. Still, even for the very large setting of  $.5M \times .5B$ , the algorithm finished in a reasonable time of less than 17 minutes.

Figure 9.  $AA^T$  sparsity on VLS2.

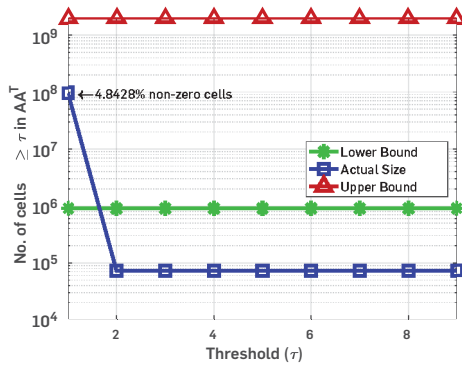


Figure 10. Scalability on  $n$  and  $m$ .

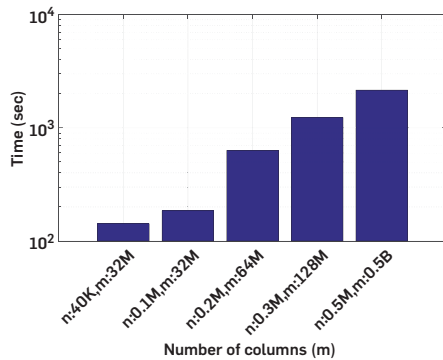
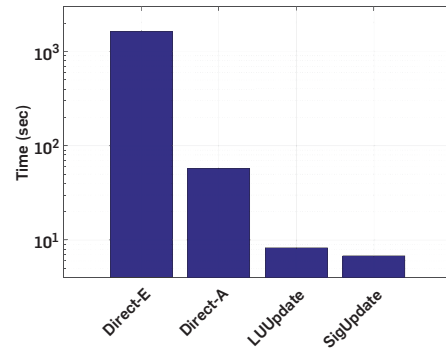


Figure 11. Dynamic-update performance on network  $p2p-3$ .



**Dynamic signal reconstruction results.** Our last set of experiments is for handling the dynamic updates. The results for dynamic scenario for  $p2p-3$  are given in Figure 11. For expounding the effects of our dynamic approach through signature matrix, we also considered adopting LU decomposition for signature matrix (LUUpdate). As is evident, both LUUpdate and SigUpdate perform well and SigUpdate slightly outperforms the other. This is because LUUpdate requires solving two systems of linear equations.

## 8. RELATED WORK


**Linear algebraic techniques for solving SRP:** There has been extensive work on solving the system of linear equations using diverse techniques such as computing the pseudoinverse of  $A$ <sup>13</sup> or performing singular value decomposition (SVD) on  $A$ , and iterative algorithms for solving the linear system.<sup>13</sup> However, none of these methods scale for large-scale signal reconstruction problems. A key bottleneck in these approaches is the computation of the pseudoinverse for matrix  $A$ . Any matrix  $B$  such that  $ABA = A$  is defined as a pseudoinverse for  $A$ . It is possible to identify “the infinitely many possible generalized inverses,”<sup>13</sup> each with its own advantages and disadvantages. Moore-Penrose Pseudoinverse (MPP)<sup>11</sup> is one of the most well-known and widely used pseudoinverse. MPP is the pseudoinverse that has the smallest Frobenius norm, minimizes the least-square fit in overdetermined systems, and finds the shortest solution in the underdetermined ones. However, none of the pseudoinverse definitions suits our purpose of finding the solution  $X$  that minimizes the  $\ell_2$ -distance from a prior. Furthermore, computing pseudoinverses is often done by SVD that is computationally very expensive.

## 9. CONCLUSION

In this paper, we investigated how a wide ranging problem of large-scale signal reconstruction can benefit from techniques developed by the database community. Efficiently solving SRP has a number of applications in diverse domains such as network traffic engineering, astronomy, and medical imaging. We propose an algorithm DIRECT based on the Lagrangian dual form of SRP. We identify a number of computational bottlenecks in DIRECT and

evaluate the use of database techniques such as sampling and similarity joins for speeding them up without much loss in accuracy. Our experiments on networks that are orders of magnitude larger than prior work show the potential of our approach.

### ACKNOWLEDGMENTS

This work was supported in part by AT&T, the National Science Foundation under grants 1343976, 1443858, 1624074, and 1760059, and the Army Research Office under grant W911NF-15-1-0020. 

### References

- Asudeh, A., Augustine, J., Nazi, A., Thirumuruganathan, S., Zhang, N., Das, G., Srivastava, D. Scalable algorithms for signal reconstruction by leveraging similarity joins. *Vldb J.* 29, 2 (2020), 681–707.
- Asudeh, A., Nazi, A., Augustine, J., Thirumuruganathan, S., Zhang, N., Das, G., Srivastava, D. Leveraging similarity joins for signal reconstruction. *PVLDB* 10, 11 (2018), 1276–1288.
- Beyer, K., Gemulla, R., Haas, P.J., Reinwald, B., Sismanis, Y. Distinct-value synopses for multiset operations. *Commun. ACM* 10, 52 (2009), 87–95.
- Broder, A.Z. On the resemblance and containment of documents. In *SEQUENCES* (1997), IEEE, 21–29.
- Chaudhuri, S., Ganti, V., Kaushik, R. A primitive operator for similarity joins in data cleaning. In *ICDE* (2006). IEEE.
- Cohen, E., Kaplan, H. Tighter estimation using bottom k sketches. *PVLDB* 1, 1 (2008), 213–224.
- Grangeat, P., Amans, J.-L. *Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, Vol. 4. Springer Science & Business Media, Springer Netherlands, 1996.
- Kalinin, S.V., Strelcov, E., Belianinov, A., Somnath, S., Vasudevan, R.K., Lingerfelt, E.J., Archibald, R.K., Chen, C., Proksch, R., Laanait, N., et al. Big, deep, and smart data in scanning probe microscopy. *ACS Nano* 10, 10 (2016), 9068–9086.
- Liu, Z., Shi, Z., Jiang, M., Zhang, J., Chen, L., Zhang, T., Liu, G. Using MC algorithm to implement 3d image reconstruction for yunnan weather radar data. *J. Comput. Commun.* 05, 5 (2017), 50–61.
- Massey, R., Rhodes, J., Ellis, R., Scoville, N., Leauthaud, A., Finoguenov, A., Capak, P., Bacon, D., Aussel, H., Kneib, J.-P., et al.

Dark matter maps reveal cosmic scaffolding. *arXiv preprint astro-ph/0701594* (2007).

- Penrose, R. A generalized inverse for matrices. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 51 (1955), 406–413.
- Trefethen, L.N., Bau III, D. Technical report. *Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics. ISBN 978-0-89871-361-9, 1997.
- Vogel, C.R. *Computational methods for inverse problems*. SIAM, 2002.
- Wiaux, Y., Jacques, L., Puy, G., Scaife, A.M., Vanderghynst, P. Compressed sensing imaging techniques for radio interferometry. *Monthly Notices of the Royal Astronomical Society* 3, 395 (2009), 1733–1742.
- Zhang, Y., Roughan, M., Duffield, N., Greenberg, A. Fast accurate computation of large-scale IP traffic matrices from link loads. In *SIGMETRICS*, Volume 31, 2003.
- Zhu, Y., Li, Z., Zhu, H., Li, M., Zhang, Q. A compressive sensing approach to urban traffic estimation with prob vehicles. *IEEE Trans. Mobile Comput.* 11, 12 (2012), 2289–2302.

**Abolfazl Asudeh** (asudeh@uic.edu), University of Illinois at Chicago.

**Azade Nazi** (azade.nazi@google.com), Google Brain.

**Jees Augustine and Gautam Das** (([jees.augustine@mavs, gdas@cse].uta.edu)), University of Texas at Arlington.

**Nan Zhang** (nzhang@american.edu), Kogod School of Business, American University.

**Saravanan Thirumuruganathan** (sthirumuruganathan@hbku.edu.qa), QCRI, HBKU.

**Divesh Srivastava** (divesh@research.att.com), AT&T Labs-Research.



Copyright held by authors/owners.

# Computing and the National Science Foundation, 1950-2016

*Building a Foundation for Modern Computing*

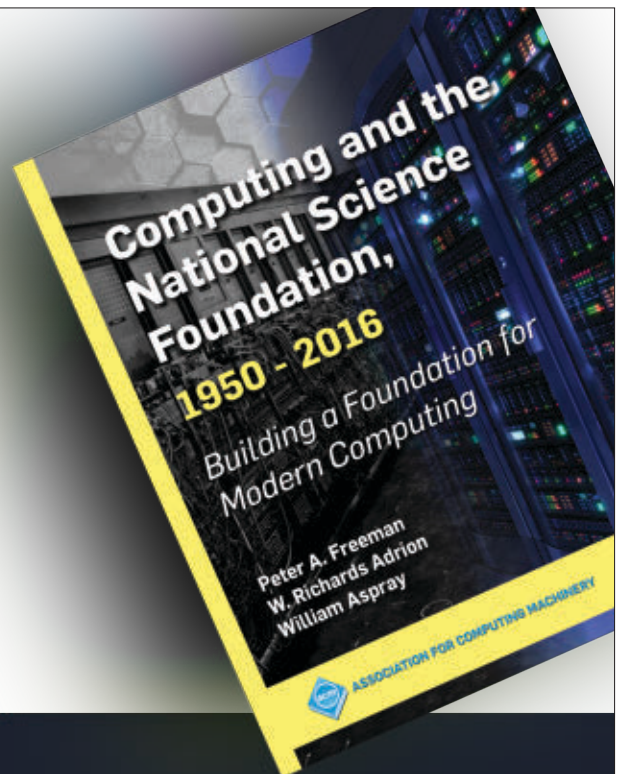
**Peter A. Freeman**  
**W. Richards Adrion**  
**William Aspray**

ISBN: 978-1-4503-7271-8

DOI: 10.1145/3335772

<http://books.acm.org>

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



**ACM BOOKS**  
Collection II



# CAREERS

## Auburn University

### Multiple Faculty Positions in Computer Science and Software Engineering

The Department of Computer Science and Software Engineering (CSSE), situated within the Samuel Ginn College of Engineering, invites applications for multiple tenure-track faculty positions. We seek candidates at the Assistant Professor level, although outstanding candidates at a senior level will also be considered. A Ph.D. degree in computer science, software engineering, or a closely related field must be completed by the start of appointment.

While applications from candidates with expertise in any area of computer science will be considered, focus areas are **Human-Computer Interaction (HCI)**, **Systems** (broadly defined to include operating systems, compilers, programming languages, software environments, advanced architectures, parallel and distributed computing, etc.), **Software Engineering (SE)**, and **Data Science**. *We are especially interested in candidates with expertise in multiple areas such as HCI & SE, HCI & Artificial Intelligence (e.g., intelligent interfaces), SE & Security, and Systems & Quantum Computing. We welcome applications from women and those belonging to underrepresented groups in computer science.*

CSSE is the highest ranked computer science department in Alabama, fifth among SEC schools, and among the top 100 departments in the nation according to the latest rankings from U.S. News and World Report. The department has 25 full-time tenure-track and 7 teaching-track faculty members, who support a dynamic research enterprise and strong undergraduate and graduate programs (M.S. in CSSE, M.S. in Cybersecurity Engineering, M.S. in Data Science & Engineering, and Ph.D. in CSSE). Current student enrollment is over 1200 undergraduate and over 200 graduate students. Further information may be found at the department's home page <http://www.eng.auburn.edu/csse>.

Auburn University is one of the nation's premier, public, Carnegie R1 research, and land-grant institutions. Auburn residents enjoy a thriving community, recognized as one of the "best small towns in America," with moderate climate and easy access to major cities or to beach and mountain recreational facilities.

Applicants should submit a cover letter, curriculum vita, research vision, teaching philosophy, and names of three to five references at <https://www.auemployment.com/postings/20004>. There is no application deadline. The application review process will begin December 1 2020 and continue until successful candidates are identified.

Selected candidates must be able to meet eligibility requirements to work legally in the United States at the time of appointment for the proposed term of employment. Auburn University is an Affirmative Action/Equal Opportunity Employer.

## University of Illinois at Chicago

### Lecturer Non-Tenure Track

The Computer Science Department at the University of Illinois at Chicago is seeking two full-time teaching faculty members. The Lecturer teaching track is a long-term career track that starts with the Lecturer position and offers opportunities for advancement to Senior Lecturer. Candidates would work alongside 16 full-time teaching faculty with over 150 years of combined teaching experience and 12 awards for excellence. The standard teaching load is 1-3 undergraduate courses per semester, depending on enrollment.

The first opening is targeted for computer ethics and technical communications. Minimum qualifications include a graduate degree in ethics and/or communications-related field. Some experience in computer science, or a related field is preferred, but not required.

Areas of interest for the second opening include introductory programming, data structures, computer organization/systems, web development, data science, software engineering, and machine learning. Minimum qualifications include an MS in Computer Science.

Candidates for either position must have either (a) demonstrated evidence of effective teaching, or (b) convincing argument of future dedication and success in the art of teaching.

The University of Illinois at Chicago (UIC) is one of the top-ten most diverse universities in the US (US News and World Report), a top-ten best value (Wall Street Journal and Times Higher Education) and a Hispanic-serving institution. Chicago epitomizes the modern, livable, vibrant city. Located on the shore of Lake Michigan, Chicago offers an outstanding array of cultural, culinary, recreational, and sporting experiences. In addition to the lakefront and theater districts, Chicago boasts one of the world's tallest and densest skylines, an 8100-acre park system, professional teams in all major sports, and extensive public transit and biking networks.

Applications are submitted online at <https://jobs.uic.edu/>. In the online application include a curriculum vitae, names and addresses of at least three references, a statement providing evidence of effective teaching, a statement describing past experience in activities that promote diversity and inclusion (or plans to make future contributions), recordings of recent teaching activities either in-person or online, and recent teaching evaluations. For additional information please contact Dr. John Bell, Committee Chair, [jbelle@uic.edu](mailto:jbelle@uic.edu).

For fullest consideration, please apply by January 3, 2021. We will continue to accept and review applications until the positions are filled.

The University of Illinois at Chicago is an Equal Opportunity, Affirmative Action employer. Minorities, women, veterans and individuals with disabilities are encouraged to apply.

Offers of employment by the University of Illinois may be subject to approval by the University's Board of Trustees and are made contingent upon the candidate's successful completion of any criminal background checks and other pre-employment assessments that may be required for the position being offered. Additional information regarding such pre-employment checks and assessments may be provided as applicable during the hiring process.

The University of Illinois System requires candidates selected for hire to disclose any documented finding of sexual misconduct or sexual harassment and to authorize inquiries to current and former employers regarding findings of sexual misconduct or sexual harassment. For more information, visit <https://www.hr.uillinois.edu/cms/One.aspx?portalId=4292&pageId=1411899>.

## University of Illinois at Chicago

### Open Rank Non-Tenure Track Teaching Faculty

The Computer Science Department at the University of Illinois at Chicago (UIC) seeks two full-time teaching faculty. The Clinical Professor teaching track is a long-term career track that starts with Clinical Assistant Professor position and offers advancement to Clinical Associate and Clinical Full Professor. Candidates would join 16 full-time teaching faculty with over 150 years of combined teaching experience and 12 awards for excellence. The standard teaching load is 1-3 undergraduate courses per semester, depending on enrollment.

The first opening is targeted for computer ethics and technical communications. Minimum qualifications include a PhD in ethics and/or communications-related field. Some experience in computer science, or a related field is preferred, but not required.

Areas of interest for the second opening include introductory programming, data structures, computer organization/systems, web development, data science, software engineering, and machine learning. Minimum qualifications include a PhD in Computer Science.

Candidates for either position must have either (a) demonstrated evidence of effective teaching, or (b) convincing argument of future dedication and success in the art of teaching. Candidates interested in Computer Science Education research are encouraged to apply.

UIC is a top-ten most diverse university in the US (US News and World Report), a top-ten best value (Wall Street Journal and Times Higher Education) and a Hispanic-serving institution. Chicago epitomizes the modern, livable, vibrant city. On the shore of Lake Michigan, Chicago offers outstanding cultural, culinary, and recreational experiences. Besides the lakefront and theater, Chicago boasts one of the world's tallest densest skylines, an 8100-acre park system, professional teams in all major sports, and extensive public transit and biking networks.

Applications are submitted online at <https://jobs.uic.edu/>. In the online application include a curriculum vitae, names and addresses of at least three references, a statement providing evidence of effective teaching, a statement describing past experience in activities that promote diversity and inclusion (or plans to make future contributions), recordings of recent teaching activities either in-person or online, and recent teaching evaluations. For additional information please contact Dr. John Bell, Committee Chair, [jbelle@uic.edu](mailto:jbelle@uic.edu).

For fullest consideration, please apply by January 3, 2021. We will continue to accept and review applications until the positions are filled.

The University of Illinois at Chicago is an Equal Opportunity, Affirmative Action employer. Minorities, women, veterans, and individuals with disabilities are encouraged to apply.

Offers of employment by the University of Illinois may be subject to approval by the University's Board of Trustees and are made contingent upon the candidate's successful completion of any criminal background checks and other pre-employment assessments that may be required for the position being offered. Additional information regarding such pre-employment checks and assessments may be provided as applicable during the hiring process.

The University of Illinois System requires candidates selected for hire to disclose any documented finding of sexual misconduct or sexual harassment and to authorize inquiries to current and former employers regarding findings of sexual misconduct or sexual harassment. For more information, visit <https://www.hr.uillinois.edu/cms/One.aspx?portalId=4292&pageId=1411899>.

### University of Maryland, Baltimore County

Assistant/Associate/Full Professor  
(Open Rank)

The Department of Information Systems (IS) at UMBC invites applications for an open rank tenure-track faculty position starting August 2021. Successful candidates will complement and extend our current strengths. Candidates with research interests cross-cutting multiple areas are particularly encouraged to apply. Candidates must have earned a PhD in related fields no later than August 2021.

Candidates are expected to establish a collaborative, externally funded, and nationally recognized research program and contribute to teaching a variety of graduate and undergraduate courses offered by the department effectively. We expect candidates to be innovative in terms of pedagogical methods, course content, and curriculum development, and be committed to advising, mentoring and supporting student success. All candidates should have experience in – or have the potential for – building an equitable and diverse scholarly environment in teaching, mentoring, research, life experiences, or service. Candidates for the Associate and Full Professor rank should also demonstrate a track record of inclusive excellence. Candidates for the Associate Professor rank should also have a strong record of research, teaching, service, and a sustained externally-funded research program. Candidates for the Full Professor rank should also demon-



## FACULTY POSITIONS

### Department of Computer Science

The Department of Computer Science at Virginia Tech is in a period of dramatic growth and opportunity. With substantial multi-year investments from the Commonwealth of Virginia and infrastructure investments by Virginia Tech, we anticipate hiring multiple faculty members at all ranks and in all areas for several years. We seek candidates motivated to contribute to a collegial, interdisciplinary community with a strong tradition of both fundamental and applied research. We embrace Virginia Tech's motto, *Ut Prosim* ("That I May Serve"): we are committed to research, education, service, and inclusivity that makes a positive difference in the lives of people, communities, and the world.

We seek candidates at all ranks and in all areas of computer science, and from all backgrounds and lived experiences. The positions include packages and resources to enable success. Our new colleagues will benefit from the department's highly-focused faculty development and mentoring program, as well as numerous successful collaborations with government, national labs, and industry partners. Candidates for all positions must have a Ph.D. in computer science or a related field at the time of appointment and a rank-appropriate record of scholarship and collaboration in computing research. Tenured and tenure-track faculty are expected to initiate and develop independent research that is internationally recognized for excellence, conscientiously mentor research-oriented graduate students, teach effectively at both graduate and undergraduate levels, and serve the university and their professional communities.

The department fully embraces Virginia Tech's commitment to increase faculty, staff, and student diversity; to ensure a welcoming, affirming, safe, and accessible campus climate; to advance our research, teaching, and service mission through inclusive excellence; and to promote sustainable transformation through institutionalized structures. We cultivate a working environment that respects differences in gender, race, ethnicity, sexual orientation, physical ability/qualities, and religious status. We strongly encourage applications from traditionally underrepresented communities to join us in this critical endeavor.

The department currently has 57 faculty members, including 47 tenured or tenure-track faculty, 15 early career awardees, and numerous recipients of faculty awards from IBM, Intel, AMD, Microsoft, Google, Facebook, and others. CS faculty members direct several interdisciplinary research centers, including the Center for Human-Computer Interaction and the Discovery Analytics Center. The department is home to over 1,200 undergraduate majors and over 300 graduate students, with university commitments to grow all programs significantly. The department is in the College of Engineering, whose undergraduate program ranks 13th and graduate program ranks 31st among all U.S. engineering schools (*USN&WR*). Virginia Tech's main campus is located in Blacksburg, VA, in an area consistently ranked among the country's best places to live. In addition, our program in the Washington, D.C., area offers unique proximity to government and industry partners and is also expanding rapidly, with Virginia Tech's exciting new Innovation Campus in Alexandria, VA, slated to open in 2024. Candidates for faculty positions at the Innovation Campus are encouraged to apply to the separate announcement for those opportunities.

The positions require occasional travel to professional meetings. Selected candidates must pass a criminal background check prior to employment. Applications must be submitted online to [jobs.vt.edu](https://jobs.vt.edu) for position 514466. Application review will begin on 11/20/20 and continue until the positions are filled. Inquiries should be directed to Dr. Ali R. Butt, search committee chair, at [facdev@cs.vt.edu](mailto:facdev@cs.vt.edu).

*Virginia Tech is an equal opportunity/affirmative action institution.  
A criminal background check is the condition of employment with Virginia Tech.*

strate leadership in their field, hold an excellent academic record, and show a history of securing external funds for multiple sizable research projects. We are particularly interested in receiving applications from individuals who are members of groups that historically have been under-represented in the professoriate.

Applications for the positions must be submitted as PDF files via Interfolio at <https://apply.interfolio.com/81030>. Review of applications will start in December 2020 but will continue until positions are filled. All interviews will be conducted online but applicants are welcome to talk to IS faculty to learn about Baltimore and the surrounding area.

Candidates' experience will be evaluated commensurate to the rank to which they are applying. For inquiries, please email to [is\\_faculty\\_search\\_2020@umbc.edu](mailto:is_faculty_search_2020@umbc.edu). An informational webinar will be also held in late November or early December. If you are interested in the webinar, please register at <https://forms.gle/CmugCMfP-dnPRoT386>. Review of applications will begin in December 2020 and will continue until the position is filled. For best consideration, please apply by January 15, 2021.

UMBC is an Affirmative Action/Equal Opportunity Employer and welcomes applications from minorities, women, veterans, and individuals with disabilities.

As an institution that receives federal financial assistance, UMBC adheres to Title IX and does not discriminate on the basis of sex.



## ADVERTISING IN CAREER OPPORTUNITIES

**How to Submit a Classified Line Ad: Send an e-mail to [acmm mediasales@acm.org](mailto:acmm mediasales@acm.org). Please include text, and indicate the issue/or issues where the ad will appear, and a contact name and number.**

**Estimates: An insertion order will then be e-mailed back to you. The ad will be typeset according to CACM guidelines. NO PROOFS can be sent. Classified line ads are NOT commissionable.**

**Deadlines: 20th of the month/2 months prior to issue date. For latest deadline info, please contact:**

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

**Career Opportunities Online: Classified and recruitment display ads receive a free duplicate listing on our website at:**

<http://jobs.acm.org>

**Ads are listed for a period of 30 days.**

**For More Information Contact:**

**ACM Media Sales  
at 212-626-0686 or  
[acmm mediasales@acm.org](mailto:acmm mediasales@acm.org)**

## ACM Computing Surveys (CSUR)

2018 JOURNAL  
IMPACT FACTOR:  
6.131

*Integration of computer science and engineering knowledge*



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)





[CONTINUED FROM P. 120] have tremendous fall-off, because the omni-directional antenna is sending out energy in all directions. When you steer that energy in a particular direction, you can get a lot of the energy back, and there are different techniques. You can use antenna designs like horn antennas, for example, to get this directivity. But the beauty of MIMO is that you use software and electronic steering techniques to dynamically point the energy exactly in the direction that you want it to go, depending on where the receiver moves. That's in theory. In practice, it's hard to do because any interference scatters the energy in all directions. Millimeter-wave is much more sensitive to interference because it requires this directional steering in order to get reasonable performance.

#### What are some of the techniques you've explored?

There are many open questions. We've done some work looking at the fundamental capacity limits of massive MIMO arrays that adapt to time-varying channels. We started with perfect conditions, where you can estimate the channel perfectly and feed it back instantaneously. Of course, that's a very idealized setting. In a typical massive MIMO setting, you need to measure the antenna gain from tens or even hundreds of antenna elements at the transmitter to every one of the antenna elements at the receiver. That's much more challenging. So we've also looked at techniques for situations where you can't do that kind of dynamic adaptation. What if you estimated the channel imperfectly—how would you deal with interference? What if you stopped trying to do any kind of channel estimate and did blind MIMO decoding? We've also looked into adapting the antenna arrays to meet the requirements of different applications, because some applications don't require such high-performance gains.

#### So you're trying to match what you're doing at the physical layer with the requirements at the application layer.

The next generation of wireless networks needs to support a much broader range of applications. The goal of each generation of cellular has always been getting to higher data rates, but

**“There's a price to be paid for machine learning, in terms of computational complexity and latency.”**

what we're looking at now are low-latency applications like autonomous driving, and networks so far have not really put hard latency constraints into their design criteria. If you exceed the latency constraints on your video or audio applications, it just means that quality is poor, or maybe the connection is dropped. That's not acceptable for a real-time autonomous vehicle application. Networks also need to be able to support soft constraints on energy consumption for low-power Internet of Things devices, which might run off a battery that can't be recharged.

#### Let's talk about machine learning, which you found can trump theory in equalizing unknown or complex channels.

I was very skeptical of jumping onto the bandwagon of machine learning, but when you don't have good models, machine learning is an interesting tool for figuring out the end-to-end optimization of a system. We first applied machine learning when we were working on molecular communication: using molecules instead of electromagnetic waves to send ones and zeros. We used an acid for one and a base for zero and sent it out through a liquid channel. The way the signal propagates is by diffusion, and there's no good channel model for that. You also need to equalize it, because the chemicals sit around in the channel for a long time. If you send a lot of ones, then the channel has too much acid in it, and when you send a base, it will get destroyed by the acid.

#### In that situation, you found that machine learning worked better than any existing techniques.

That's right. Later, we started looking at machine learning more broadly

for channel equalization on traditional wireless channels. The optimal technique is the Viterbi algorithm, and we found that it can't be beat under ideal conditions, where you know the channel perfectly and you have no complexity constraints. But when you relax those perfect assumptions, it turns out that machine learning can do better.

Of course, it's not always the case that you should go to machine learning as soon as you move away from perfect conditions. There's a price to be paid for machine learning, in terms of computational complexity and latency. But to me, the meta-lesson is that having domain knowledge plus some knowledge of machine learning is much more valuable to solving domain-specific problems than having very deep knowledge of machine learning, but not really understanding the specific problem you're trying to solve. We understood the problem of equalization well, so we were able to take this tool and use it very efficiently to reach a solution.

#### You were recently appointed dean of Princeton University's School of Engineering and Applied Science. What are some of your goals?

Princeton already has a strong group of people who are working on wireless communication and networking. For my own research, I'm excited to work with these Princeton colleagues, as well as researchers in nearby wireless groups at NYU and Rutgers. There's been a resurgence of interest in wireless lately, and in bridging the digital divide in the pandemic, so it's a very exciting time to be working in the field.

I join Princeton at a time when it is growing the size of its engineering faculty by almost 50%, building an entirely new neighborhood with new buildings for all its engineering departments and interdisciplinary institutes, and also building a separate part of campus dedicated to innovation, entrepreneurship, and forging stronger ties with industry. I'm really excited to be the incoming dean at such a transformational time for Princeton Engineering.

Leah Hoffmann is a technology writer based in Piermont, NY, USA.

© 2021 ACM 0001-0782/21/2 \$15.00

## Q&amp;A

# Bringing Stability to Wireless Connections

*2020 Marconi Prize recipient Andrea Goldsmith on MIMO technologies, millimeter-wave communications, and her goals as the new dean of Princeton University's School of Engineering and Applied Science.*

COMMUNICATION IS MORE important than ever, with everything from college to CrossFit going virtual during the COVID-19 pandemic. Nobody understands this better than 2020 Marconi Prize recipient Andrea Goldsmith, who has spent her career making the wireless connections on which we rely more capable and stable. A pioneer of both theoretical and practical advances in adaptive wireless communications, Goldsmith spoke about her work on multiple-input and multiple-output (MIMO) channel performance limits, her new role as the incoming dean at Princeton University's School of Engineering and Applied Science, and what's next for networking.

**As an undergrad, you studied engineering at the University of California, Berkeley. What drew you to wireless communications?**

After I got my undergraduate degree, I went to work for a small defense communications startup. It was a great opportunity, because I was working on really hard problems with people who had advanced degrees. We were looking at satellite communication systems and antenna array technology. I was really motivated to go back to graduate school because I wanted to learn more.

**This was around the time that commercial wireless was starting to take off; cellular systems in particular.**

By the time I went to graduate school, in 1989, they were starting to talk about



second-generation cellular standards. There was a big debate about what the technology should be. I found that whole area fascinating, and it's what I ended up focusing on initially.

**Later, after joining Stanford's Electrical Engineering department, you made groundbreaking advances in multiple-input and multiple-output (MIMO) channel performance limits.**

We had looked at direction-finding techniques at the defense communications startup in the '80s, which exposed me to the MUSIC and ESPRIT algorithms for direction-finding with multiple antennas. During graduate school, I spent two

summers working at AT&T Bell laboratories with Gerry Foschini, whose work informed a lot of the early MIMO techniques, following up on the groundbreaking work of A. Paulraj at Stanford.

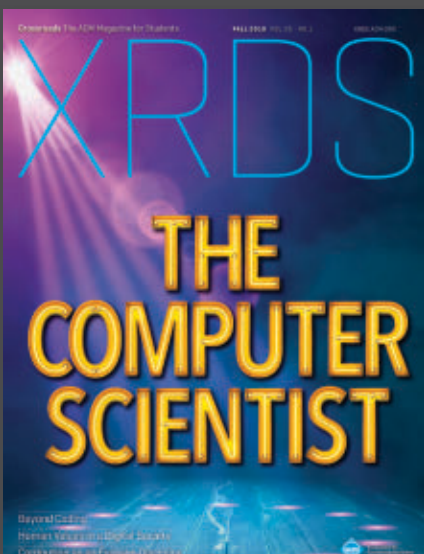
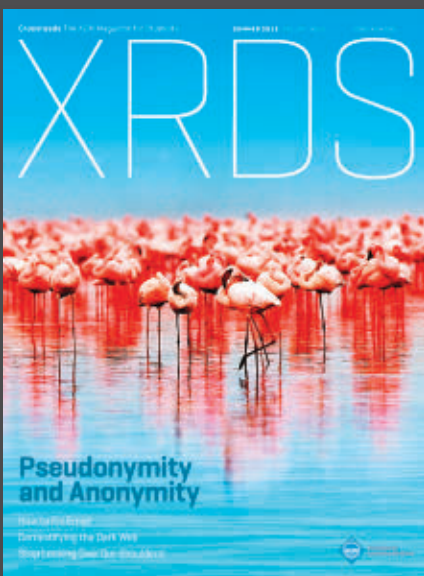
A few years after I came to Stanford, MIMO technology emerged as a really compelling one for capacity gain. So my group started looking at how to handle the dynamic adaptation of multiple antenna systems. We'd been working on dynamic adaptation of single antenna systems, and that was a natural area to expand into.

**More recently, you've begun to explore deployments in the millimeter wave band, and in particular, millimeter-wave massive MIMO technologies. Can you talk about your work in that area?**

Millimeter-wave is a really interesting spectral band to explore for commercial wireless. The biggest attraction is the amount of spectrum that's available—tens of gigahertz of spectrum. We have to find ways to utilize that, especially given how much of the lower bands are already occupied.

**But millimeter-wave communication is challenging even at relatively short ranges, because it's very inefficient.**

If you have a single, omni-directional antenna, the power falls off relative to one over the frequency squared. So when you go up to these very high frequencies, you [CONTINUED ON P. 119]



# XRDS

At *XRDS*, our mission is to empower computer science students around the world. We deliver high-quality content that makes the complexity and diversity of this ever-evolving field accessible. We are a student magazine run by students, for students, which gives us a unique opportunity to share our voices and shape the future leaders of our field.

**Accessible, High-Quality, In-Depth Content** We are dedicated to making cutting-edge research within the broader field of computer science accessible to students of all levels. We bring fresh perspectives on core topics, adding socially and culturally relevant dimensions to the lessons learned in the classroom.

**Independently Run by Students** *XRDS* is run as a student venture within the ACM by a diverse and inclusive team of engaged student volunteers from all over the world. We have the privilege and the responsibility of representing diverse and critical perspectives on computing technology. Our independence and willingness to take risks make us truly unique as a magazine. This serves as our guide for the topics we pursue and in the editorial positions that we take.

**Supporting and Connecting Students** At *XRDS*, our goal is to help students reach their potential by providing access to resources and connecting them to the global computer science community. Through our content, we help students deepen their understanding of the field, advance their education and careers, and become better citizens within their respective communities.

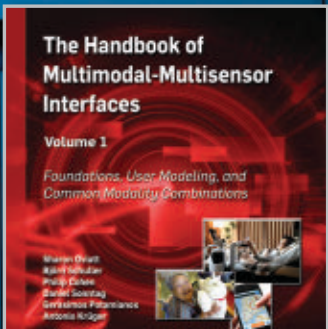
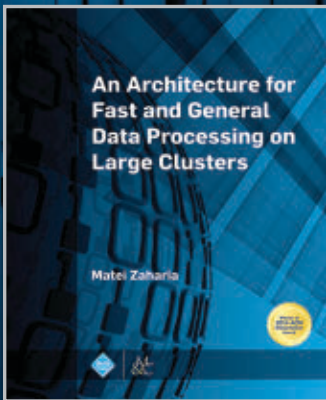
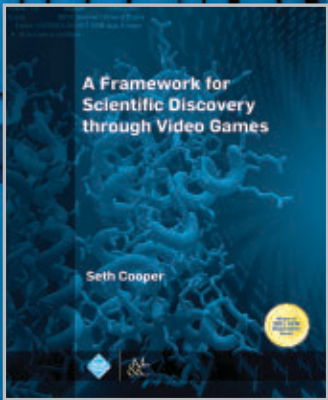
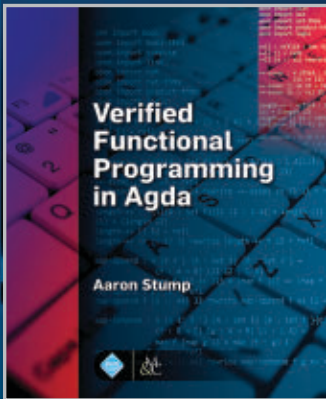
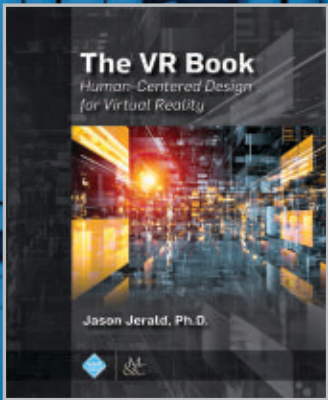
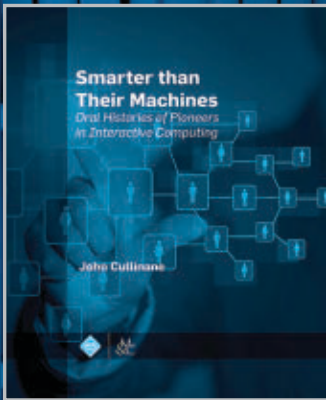
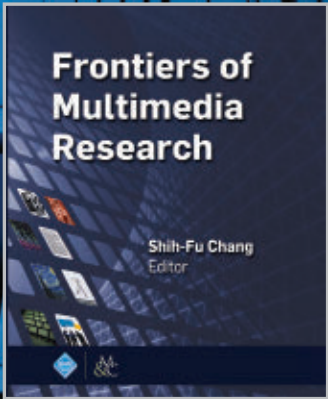
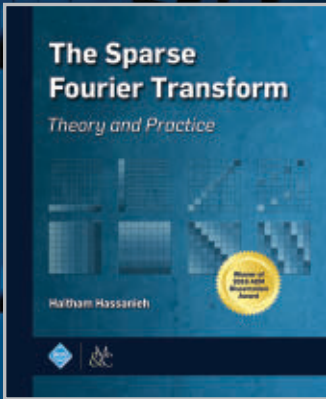
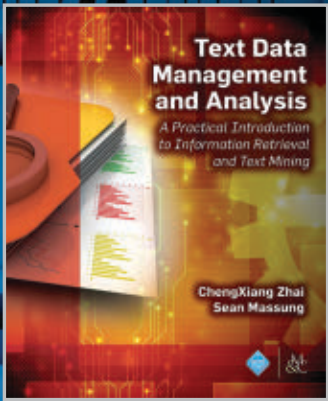
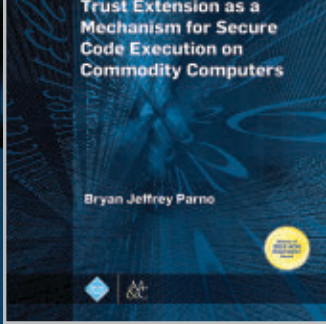
*XRDS* is the flagship magazine for student members of the Association for Computing Machinery [ACM].

[www.xrds.acm.org](http://www.xrds.acm.org)



Association for  
Computing Machinery





# 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)