

COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

04/2019 VOL.62 NO.04

Special Section on Europe Region



Cybersecurity in
the Quantum Era

Metrics That Matter

Analytics for Managerial Work

The Future of Data Storage

Association for
Computing Machinery



The Pragmatic Wisdom of Michael Stonebraker

Making Databases Work

This book celebrates Michael Stonebraker's accomplishments that led to his 2014 ACM A.M. Turing Award "for fundamental contributions to the concepts and practices underlying modern database systems."

The book describes, for the broad computing community, the unique nature, significance, and impact of Mike's achievements in advancing modern database systems over more than forty years. Today, data is considered the world's most valuable resource, whether it is in the tens of millions of databases used to manage the world's businesses and governments, in the billions of databases in our smartphones and watches, or residing elsewhere, as yet unmanaged, awaiting the elusive next generation of database systems. Every one of the millions or billions of databases includes features that are celebrated by the 2014 Turing Award and are described in this book.

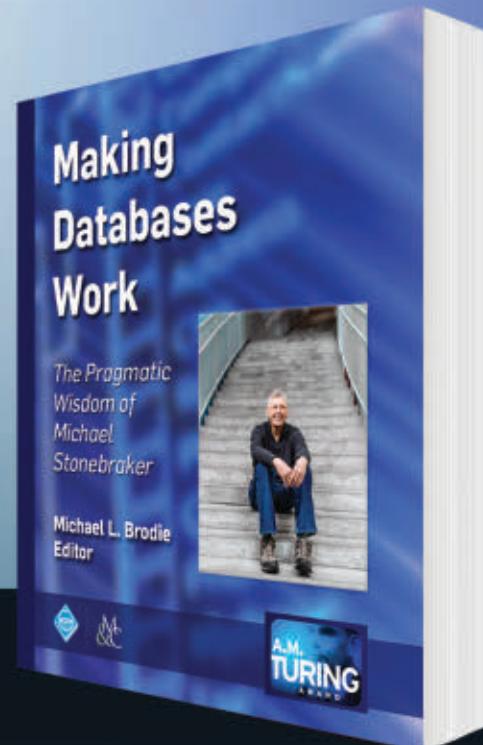
Edited by Michael L. Brodie

ISBN: 978-1-94748-719-2

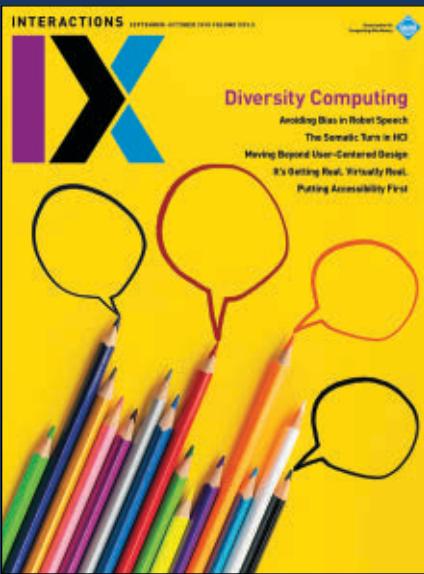
DOI: 10.1145/3226595

<http://books.acm.org>

<http://www.morganclaypoolpublishers.com/acm>



ACM BOOKS



Association for
Computing Machinery



ACM Seeks New Editor(s)-in-Chief for *ACM Interactions*

The ACM Publications Board is seeking a volunteer editor-in-chief or co-editors-in-chief for its bimonthly magazine *ACM Interactions*.

ACM Interactions is a publication of great influence in the fields that envelop the study of people and computers. Every issue presents an array of thought-provoking commentaries from luminaries in the field together with a diverse collection of articles that examine current research and practices under the HCI umbrella.

For more about *ACM Interactions*, see <http://interactions.acm.org>

Job Description The editor-in-chief is a volunteer position responsible for organizing all editorial content for every issue. These responsibilities include: proposing articles to prospective authors; overseeing the magazine's editorial board and contributors; creating new editorial features, columns, and much more.

An annual stipend will be available for the hiring of an editorial assistant.

Financial support will also be provided for any travel expenses related to this position.

Eligibility Requirements The EiC search is open to applicants worldwide. Experience in and knowledge about the issues, challenges, and advances in human-computer interaction is a must.

The ACM Publications Board has convened a special search committee to review all candidates for this position.

Please send your CV and vision statement of 1,000 words or less expressing the reasons for your interest in the position and your goals for *Interactions* to the search committee at eicsearch@interactions.acm.org, Subject line: RE: Interactions.

**The deadline for submissions is June 1, 2019 or until position is filled.
The editorship will commence on October 1, 2019.
You must be willing and able to make a three-year commitment to this post.**

COMMUNICATIONS OF THE ACM

Departments

- 5 **Cerf's Up**
In Debt to the NSF
By Vinton G. Cerf

- 6 **BLOG@CACM**
Pondering Variables and Direct Instruction

Robin K. Hill considers the nature of variables, while Mark Guzdial reflects on renewed interest in the “direct instruction model.”

- 21 **Calendar**

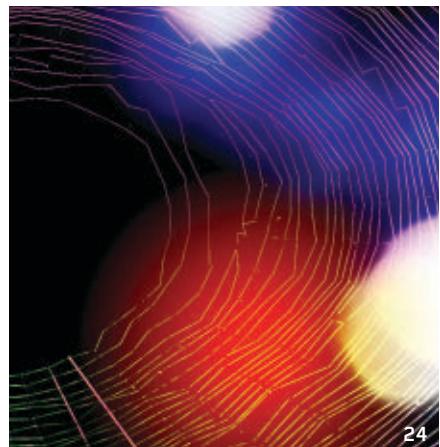
Last Byte

- 144 **Upstart Puzzles**
Fighting for Lava
By Dennis Shasha

News



Viewpoints



- 9 **Soft Robots Look to New Environments**
These non-standard automata appear best suited for applications under water and in space.
By Chris Edwards

- 12 **The Future of Data Storage**
Research into next-generation storage techniques marches forward, yet tape remains the most viable, dependable medium.
By Samuel Greengard

- 15 **The Fine Line Between Coercion and Care**
Employers monitoring their workforce must balance productivity and security considerations with employee privacy and welfare concerns.
By Sarah Underwood

- 18 **Technology Strategy and Management**
Free Trade in a Digital World
Considering the possible implications for free trade, traditionally based on non-digital goods, for a modern global economy that is increasingly based on intangible products and services enabled by digital technologies.
By Mari Sako

- 22 **Kode Vicious**
Know Your Algorithms
Stop using hardware to solve software problems.
By George V. Neville-Neil

- 24 **Viewpoint**
The Web Is Missing an Essential Part of Infrastructure: An Open Web Index
A proposal for building an index of the Web that separates the infrastructure part of the search engine—the index—from the services part that will form the basis for myriad search engines and other services utilizing Web data on top of a public infrastructure open to everyone.
By Dirk Lewandowski



Association for Computing Machinery
Advancing Computing as a Science & Profession

Special Section: Europe Region

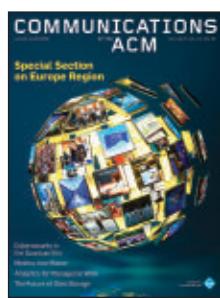
28

- 28 The special section in this issue spotlights the Europe Region with a series of articles depicting many new technologies, computing initiatives, workforce challenges, educational models, and future research directions planned for this vibrant part of the world. The cover image offers a collective glimpse of the stories inside.



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

IMAGES: (L) SPOOKY POKA AT DEBUT ART; (C) RICHARD DEVONSHIRE; (R) PNKEYES



About the Cover:
A spherical mosaic of some of the images, influences, and technologies depicted in this issue's special section on the Europe Region.
Cover illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: ELLIS group photo courtesy of ELLIS Society; Horizon Europe photo courtesy of Press EUROCHAMBRES/Flickr (CC BY 2.0); Ireland DPC image by Jarretera/Shutterstock.com; Shazam app image by XanderSt/Shutterstock.com; EU flag photo by Alexandros Michalidis/Shutterstock.com; lecture photo by Michal Cervenansky/Shutterstock.com; cobble street photo by Radikofka/Shutterstock.com; Zalando photo by nitpicker/Shutterstock.com. Additional stock images from Shutterstock.com.

Practice

80

80 Identity by Any Other Name

The complex cacophony of intertwined systems.

By Pat Helland

88 Metrics That Matter

Critical but oft-neglected service metrics that every SRE and product owner should care about.

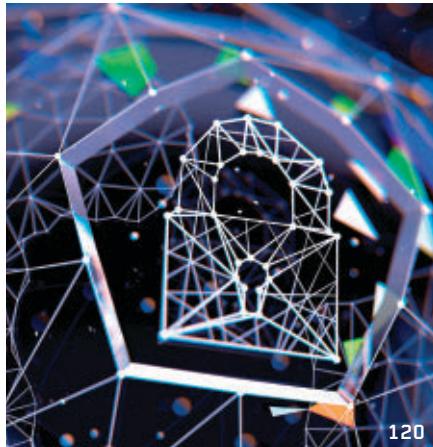
By Benjamin Treynor Sloss,
Shylaja Nukala, and Vivek Rau

95 Research for Practice:**Edge Computing**

Scaling resources within multiple administrative domains.

By Nitesh Mor

Articles' development led by queue.acm.org

Review Articles

120

110 Neural Algorithms and Computing Beyond Moore's Law

Advances in neurotechnologies are reigniting opportunities to bring neural computation insights into broader computing applications.

By James B. Aímone

120 Cyber Security in the Quantum Era

Quantum systems will significantly affect the field of cyber security research.

By Petros Wallden and Elham Kashefi



Watch the co-organizers discuss this section in the exclusive *Communications* video.
<https://cacm.acm.org/videos/cyber-security-in-the-quantum-era>

Research Highlights**132 Technical Perspective**

Was Edgar Allan Poe Wrong After All?
By Gilles Brassard

133 Fully Device Independent Quantum Key Distribution

By Umesh Vazirani and Thomas Vidick

Contributed Articles**100 Analytics for Managerial Work**

Work in finance, marketing, human resources, and operations increasingly relies on analytics—with more to come.

By Vijay Khatri and Binny M. Samuel



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.

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 Randin
Director, Office of SIG Services
Donna Cappo
Director, Office of Publications
Scott E. Delman

ACM COUNCIL

President
Cherri M. Pancake
Vice-President
Elizabeth Churchill
Secretary/Treasurer
Yannis Ioannidis
Past President
Alexander L. Wolf
Chair, SGB Board
Jeff Jortner
Co-Chairs, Publications Board
Jack Davidson and Joseph Konstan
Members-at-Large
Gabriele Anderst-Kotis; Susan Dumais;
Renée McCauley; Claudia Bauzer Medeiros;
Elizabeth D. Mynatt; Pamela Samuelson;
Theo Schlossnagle; Eugene H. Spafford
SGB Council Representatives
Sarita Adve; 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; Joseph Konstan
Board Members
Phoebe Ayers; Edward A. Fox; Chris Hankin;
Xiang-Yang Li; Nenad Medvidovic;
Sue Moon; Michael L. Nelson;
Sharon Oviatt; Eugene H. Spafford;
Stephen N. Spencer; Divesh Srivastava;
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; acmopo@acm.org

Computer Science Teachers Association
Jake Baskin
Executive Director

STAFF

DIRECTOR OF PUBLICATIONS

Scott E. Delman
caca.publisher@acm.acm.org

Executive Editor

Diane Crawford

Managing Editor

Thomas E. Lambert

Senior Editor

Andrew Rosenblom

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 Balaquit

Production Manager

Bernadette Shade

Intellectual Property Rights Coordinator

Barbara Ryan

Advertising Sales Account Manager

Ilia Rodriguez

Columnists

David Anderson; Michael Cusumano;
Peter J. Denning; Mark Guzdial;
Thomas Haigh; Leah Hoffmann; Mari Sako;
Pamela Samuelson; Marshall Van Alstyne

CONTACT POINTS

Copyright permission

permissions@hq.acm.org

Calendar items

calendar@acm.acm.org

Change of address

acmhelp@acm.org

Letters to the Editor

letters@acm.acm.org

WEBSITE

<http://acm.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

2 Penn Plaza, Suite 701, New York, NY
10121-0701
T (212) 626-0686
F (212) 869-0481

Advertising Sales Account Manager

Ilia Rodriguez

ilia.rodriguez@hq.acm.org

Media Kit

acmmEDIASales@acm.org

Association for Computing Machinery (ACM)

2 Penn Plaza, Suite 701
New York, NY 10121-0701 USA
T (212) 869-7440; F (212) 869-0481

EDITORIAL BOARD

EDITOR-IN-CHIEF

Andrew A. Chien

eic@acm.acm.org

Deputy to the Editor-in-Chief

Lihan Chen

caca.deputy.eic@gmail.com

SENIOR EDITOR

Moshe Y. Vardi

NEWS

Co-Chairs

Marc Snir and Alain Chesnais

Board Members

Monica Divitini; Mei Kobayashi;
Rajeev Rastogi; François Sillion

VIEWPOINTS

Co-Chairs

Tim Finin; Susanne E. Hambrusch;

John Leslie King; Paul Rosenbloom

Board Members

Michael L. Best; Judith Bishop;
James Grimmelmann; Mark Guzdial;

Haym B. Hirsch; 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;

Nicola 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

William Aiello; Robert Austin; Kim Bruce;

Alan Bundy; Peter Buneman; Jeff Chase;

Andrew W. Cross; Carl Gutwin;

Yannis Ioannidis; Gal A. Kaminka;

Ashish Kapoor; Igor Markov;

Lionel M. Ni; Adrian Perrig; Doina Precup;

Marie-Christine Rousset; Krishnan Sabnani;

m.c. schraefel; Ron Shamir; Alex Smola;

Sebastian Uchitel; Hannes Werthner;

Reinhard Wilhelm

RESEARCH HIGHLIGHTS

Co-Chairs

Azer Bestavros and Shriram Krishnamurthi

Board Members

Martin Abadi; Amr El Abbadi;

Animashree Anandkumar; Sanjeev Arora;

Michael Backes; Maria-Florina Balcan;

David Brooks; Stuart K. Card; Jon Crowcroft;

Alexei Efros; Bryan Ford; Alon Halevy;

Gernot Heiser; Takeo Igarashi; Sven Koenig;

Greg Morrisett; Tim Roughgarden;

Guy Steele, Jr.; Robert Williamson;

Margaret H. Wright; Nicholai Zeldovich;

Andreas Zeller

SPECIAL SECTIONS

Co-Chairs

Sriram Rajamani and Jakob Rehof

Board Members

Tao Xie; Kenjiro Taura; David Padua

ACM Copyright Notice

Copyright © 2019 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, 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals postage paid at New York, NY 10001, and other mailing offices.

POSTMASTER

Please send address changes to *Communications of the ACM*, 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA

Printed in the USA.



Association for Computing Machinery





DOI:10.1145/3313989

Vinton G. Cerf

In Debt to the NSF

THE DEFENSE ADVANCED Research Projects Agency launched the ARPANET project in 1968 and the Internet project in 1973.

In 1980, the National Science Foundation (NSF) sponsored the development of CSNET^a to connect a number of computer science departments together that had not already been connected to the ARPANET. Using a mix of dial-up phone connections, public X.25 packet network services, and access to the ARPANET, the CSNET was the programmatic forerunner of the National Science Foundation Network (NSFNET).

The NSFNET project had an enormous influence on the evolution of the Internet. The 1985 NSFNET connected five NSF-sponsored supercomputer centers together in a 56 kilobit/second network.^b A critical and controversial decision made by NSF was to use the TCP/IP protocols for the NSFNET backbone. Then, the International Organization for Standardization's Open Systems Interconnection (OSI) protocols^c were widely thought to be the direction for international computer networking. By 1987, this network^d had become congested and NSF began a new 1.5 megabit/second development in 1988 through a consortium led by MERIT,^e IBM^f and MCI,^g which would later yield operation to the non-profit organization, Advanced Networks and Services^h (ANS).

- a <https://en.wikipedia.org/wiki/CSNET>
- b https://en.wikipedia.org/wiki/Fuzzball_router
- c https://en.wikipedia.org/wiki/OSI_model
- d https://en.wikipedia.org/wiki/National_Science_Foundation_Network
- e www.merit.edu
- f www.ibm.com
- g https://en.wikipedia.org/wiki/MCI_Communications
- h https://en.wikipedia.org/wiki/Advanced_Network_and_Services

Taking advantage of the multi-network architecture of the Internet, NSF underwrote the creation of eight regional networksⁱ that would service some 4,000 universities in the U.S. and connect them to the NSFNET backbone. In addition, NSF initiated an International Connections program to underwrite the cost of international links between NSFNET and other research networks around the world.

Also in 1988, the U.S. Federal Research Internet Coordinating Council approved the interconnection of the commercial MCI Mail system^j with the NSFNET, allowing commercial traffic to flow on the U.S. Government-sponsored backbones, including ARPANET, NSFNET, the Department of Energy's ESNET^k and NASA's NSINET.^l Federal Internet Exchange (FIX) points were created to link these networks to one another.

By 1989, three commercial Internet backbone suppliers emerged: UUNET,^m Performance Systems International (PSINETⁿ) and the California Education and Research Federation Network^o (CERFnet) and they were interconnected to each other by their Commercial Internet Exchange analog of the Federal versions and were also connected to the NSFNET.

Traffic grew dramatically and the NSFNET backbone was upgraded to 45Mb/s in 1992. The next year, the High Performance Computing and High Performance Networking Ap-

plications Act of 1993^p was passed which incorporated provision for commercial traffic transiting the U.S. Government backbones. By 1994, it was becoming apparent that commercial networking was rapidly developing. NSF sponsored the creation of four Network Access Points^q (NAP) setting the stage for the termination of the NSFNET by interconnecting all the intermediate level networks to the NAPs and thus remaining interconnected to each other. The notion of NAPs evolved to become Internet eXchange Points (IXP) where many component networks could interconnect with each other and the larger Internet.

In 1995, the NSFNET backbone was retired. Its network research support functions were taken over by the Very High Speed Broadband Network Service (vBNS) operated by MCI and by Internet2, an academic consortium.^r NSF continues its vigorous support for network research to this day, spinning off new technologies leading to new commercial networking offerings. We collectively owe much to the foresight and nuanced decisions taken by the leadership of NSF's Computer, Information Systems and Engineering Directorate^s (CISE) and its Division of Computer and Network Systems. □

p <https://www.ibiblio.org/nni/boucher.html>

q one in New York operated by Sprint, one in Washington, D.C. operated by MFS, one in Chicago operated by Ameritech, and one in California operated by Pacific Bell.

r <https://www.internet2.edu/>

s <https://www.nsf.gov/dir/index.jsp?org=CISE>

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



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:10.1145/3311717

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

Pondering Variables and Direct Instruction

Robin K. Hill considers the nature of variables, while Mark Guzdial reflects on renewed interest in the “direct instruction model.”



Robin K. Hill
What Is a Variable?
<http://bit.ly/2D9XQri>
January 31, 2019

What is a variable?—the name of a value. What is an attribute?—the name of a property. These are the working definitions that most of us find adequate in daily life and daily computer science. To consider whether there is more to it is to consider an ontology, the ontology of the variable. Is there a good comprehensive definition of “variable” for students and laypersons?

First, however, let’s address the subject that arises prominently where variables meet worldly ontology—the professional design of an ontology for some real thing, in some domain, driven by a commercial need to capture some enterprise in a database. The question is, “What do we need to keep track of?”

To answer that question—there’s money in it—requires extensive knowledge of some real-world arena of activity, such as sewing, or collateralized debt obligations, or military deployments, or Bordeaux wines,⁴ along with

the skills to represent that knowledge in some data language. Working under the entity-relationship model, an expert would elicit and record the entities, attributes, and relationships that need to be exposed in the domain.

The principles of working ontology is a subject taken up by several philosophers of computing. The goal is often knowledge representation and sharing, and the objective is often organizational. Barry Smith and fellow researchers combine the abstract or formal with the domain-specific.^{6,7} Basic Formal Ontology provides templates and tools as in engineering, and addresses philosophical questions arising from those practices.¹

That’s not what we’re talking about. The question of interest here—there’s no money in it—requires probing into the nature of that workhorse abstraction, the variable. We look not to the fine-grained level of a “distinction that makes a difference” (attributed to Donald McKay³), but to the level of identifiers with associated values. Quine says that “to be is to be the value of a bounded variable,” an intriguing notion, but not the informal explication we seek. We want to articulate—toward a com-

prehensive definition—the ontology of a variable as it is used in computer programming. We mutter, “What’s the variable for the tax rate?” or “That variable has a misleading name” or “We don’t need a variable for month, because that’s a function of the date.” The ontology of something that is used so relentlessly in computing must have something to tell us about computing.

Let’s consider several questions that may provide direction. (In fact, questions are as far as we go here.)

► If a variable is the name of a value, what about a constant? The distinction is easy to explain in programming, but not so easy outside of it. Is const a special case of var, where some extra property, “mutability,” attaches to the var object? And what about the attribute of database design; what is that? Does attr break down into var and const? William Kent raised this question of the ontology of the attribute in 1978, noted its peskiness, and concluded there is really no difference between an attribute and a relationship.²

► What is the difference between a variable and a value? Is the variable a thing or some property of a thing? To say we don’t need a variable for the

month is to say the value (but not the variable) can be derived elsewhere. Because variables “go away” when values are substituted, a common view of a variable is as a “placeholder.” What is it holding a place for? For a predicate? For a proposition? For an entity? In other words, what is the shape or category of a variable? Is it an attribute of something, a whole statement about something, or the something itself? Again, in a program (or in its formalization in denotational semantics), this is clear; the difference between variable and value is embedded firmly in the process.

► What about the multiple instantiations, the fact that all occurrences of a variable are bound to the same value? To say “that variable has a misleading name” is to condemn several substrings of code. How would we include that aspect (also straightforward in a formal semantics) in a comprehensive definition?

► Is a variable something we know or something we don’t know? An algebraic definition of a variable is as an “unknown.” Yet we know a lot about a variable x that appears in an equation such as $f(x) = ax^2 + bx + c$. We know that it is an object suitable for such a setting, that is, a number, and that it belongs right there, and that its value is constrained by the given relations to other values. What’s unknown is which number. We always “know” the variable in the rather weak sense that we grasp its need and purpose. In a program, a variable’s value is something that we don’t know in the source code, the human product. In the execution, as soon as it’s bound, the value is “known” in some sense, but to what or whom?

The art of definition has a philosophy of its own.⁵ Experts in that realm would not be surprised to encounter difficulties, but computer scientists might be surprised at the subtle and complex issues that come from a mundane and familiar idea.

References

1. Buffalo Ontology Site. State University of New York at Buffalo <http://ontology.buffalo.edu/>
2. Kent, W. *Data and Reality*, North-Holland Publishing Company, Chapter 5.0, Attributes.
3. McKay, D. *Information, Mechanism and Meaning*, MIT Press, 1969
4. Noy, N.F., and McGuinness, D.L. *Ontology Development 101: A Guide to Creating Your First Ontology* <https://stanford.io/2HTnFLl>
5. Gupta, A. Definitions, *The Stanford Encyclopedia of*

- Philosophy, Edward N. Zalta, editor. <https://stanford.io/2D5Qhls>
- Smith, B. Beyond concepts: ontology as reality representation. In *Proceedings of the Third International Conference on Formal Ontology in Information Systems (FOIS 2004)*, IOS Press, pp. 73-84.
- Smith, B. Ontology (science). Formal Ontology in Information Systems, Eschenbach, C. and Grüninger, M., eds. In *Proceedings of the Fifth International Conference on Formal Ontology in Information Systems (FOIS 2008)*. IOS Press, pp. 21-35. <http://bit.ly/2TBynOB>, DOI:10.3233/978-1-58603-923-3-21



Mark Guzdial
Direct Instruction Is Better Than Discovery, But What Should We Be Directly Instructing?
<http://bit.ly/2ARuUTA>

November 7, 2018

I recently discovered Feliënne Hermans’ blog. In her latest post, she talks about the research around direct instruction and how it relates to programming. The research evidence is growing that students learn better through direct instruction rather than through a discovery-based method, where we expect students to figure things out for themselves.

Direct instruction is hot! Whereas in the 1990s we heard a lot about discovering learning, we are now slowly seeing a renewed interest in the ‘direct instruction model’ in the Netherlands. Both in language and in mathematics there is a new interest in rote practice of knowledge (“stampen,” <http://bit.ly/2D5pRAa>). As that is not a surprise, since research keeps showing that direct instruction—explanation followed by a lot of focused practice—works well (<https://nyti.ms/2HQbd5n>). It not only works well, it also equalizes: it does not matter what knowledge children already have (received at home), everyone has equal chances to acquire the basic knowledge. That is why research also shows that direct instruction works especially well for weaker pupils (<http://bit.ly/1BASeOh>).

Her point resonates strongly with me. I argued here in Blog@CACM last year (<http://bit.ly/2RzDDQR>) that we should reduce the amount that we tell students to just “figure out” in CS classes. We should teach students directly (Hermans’ point), and reduce (at least in the first classes) the amount of design and problem-solving we ask students to do.

But what should we be teaching directly? The obvious answer is “the programming language.” But there’s

a good bit of evidence that students do not learn the syntax and semantics of programming directly. Kelly Rivers had a paper at ICER 2016 (<http://bit.ly/2BhtuDa>) that I found fascinating. She studied how students learned programming structures, and found that they *didn’t*. They actually made more errors with FOR loops over the course of the semester. Of course, the students were learning, but you can’t measure their learning in terms of syntax and semantics of programming. They’re learning something else.

Elliot Soloway and Jim Spohrer argued back in the 1980s that students don’t learn programming in terms of syntax and semantics (see the CHI 1986 paper at <http://bit.ly/1BASeOh>). Rather, students were learning *plans*, useful chunks of code. Within those chunks were programming statements, but students were learning that *set* of statements, not the individual statements. Maybe we should be teaching those plans?

John Sweller argues we should be teaching programming with lots of worked examples. He argues that too much emphasis on problem-solving leads to increased cognitive load, which interferes with learning. Worked examples are completely worked out programs that students study, typically including questions that students answer about the program. The worked-example effect (<http://bit.ly/2MPGjZJ>) describes how worked examples lead to better learning than more problem-solving. For Sweller, worked examples are part of effective direct instruction for programming.

I am a follower of the research Hermans is citing in her blog post, and agree that direct instruction is better than discovery learning for introductory courses. We are still figuring out what direct instruction means in learning to program. I do not think it is about programming languages. I expect that it is more about plans, and that the methods should involve worked examples.

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. **Mark Guzdial** is a professor in the Computer Science & Engineering Division of the Engineering Education Research program at the University of Michigan.

Publish Your Work Open Access With ACM!



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

Science | DOI:10.1145/3311719

Chris Edwards

Soft Robots Look to New Environments

These non-standard automations appear best suited for applications under water and in space.

IN A LABORATORY at Yale University, a soft toy horse with prosthetic coverings around its foam-stuffed legs has taken its first tentative steps. Despite its stiff and not entirely coordinated gait, the toy demonstration may point the way toward helping space agencies put lighter, more versatile robots into space.

Rebecca Kramer-Bottiglio, assistant professor at the Yale School of Engineering & Applied Science, says she was wrestling with the problem of how to allow robots to handle a wider variety of jobs than current approaches, which often focus on performing a single function well, when the U.S. National Aeronautics and Space Administration (NASA) issued a request for novel robot designs based on lighter, plastic approaches.

Rather than attempt to lift many single-task robots into orbit, the space agency wants a single reconfigurable machine to be able to handle different tasks and, occasionally, to act as prosthetics for human astronauts. “You may need to make an exploratory locomotion robot that can go out and collect data from an unknown environment. At the same time, you may need suits to promote



"Robotic Skins" technology developed by Rebecca Kramer-Bottiglio and colleagues at the Yale School of Engineering & Applied Science enables novel designs for robots that can move more freely.

blood flow in the astronauts who are onboard,” she says. And NASA wants to avoid the weight of bulky, metal-framed robots.

“The idea I had was to have a robotic skin,” she says.

Armed with sensors and pneumatic actuators, the artificial skin can

have its function changed by changing how it is wrapped around a flexible core. With actuators aligned along the length of a flexible rod of foam or an inflatable, they can force a bending motion. Two or three close together become grippers. Rotate them 90°, and the actuators let the rod move

around like a worm. The smart skin may massage the legs of an astronaut, or act as exoskeletons to help with resistance exercises; the function changes as the skin is peeled off, rotated, and replaced.

Kramer-Bottiglio's ultimate hope is that with a sufficiently malleable interior, the skins could have two types of actuator. One would mold the interior to form appendages, while another type would move those appendages around; in effect, forming a robot that morphs based on the job it is asked to perform. "It is a vision that we are quite far from today," she concedes.

Like other researchers into robotics, Kramer-Bottiglio faces two key problems: force and control. Although the soft robots are made of lighter materials than traditional robots and should be easier to move, it is difficult to deliver large amounts of power to the artificial muscles. Engineers are still many years from being able to emulate the high power-weight ratio of organic musculature. Yale's robotized toy horse makes slow progress because the pneumatic actuators find it hard to bend its foam-

A shift away from traditional electronic robot design and construction could liberate soft automatons from their tethers and help them move more freely.

stuffed legs, and the use of open-loop control leads to motion that is far less coordinated than that of a real horse. Not only that, most of these robots need to be tethered to electronic and pneumatic or hydraulic power sources, which limits their freedom.

A shift away from traditional electronic robot design and construction could liberate soft automatons from their tethers and help them move more

freely. Several years ago, Jennifer Lewis and colleagues at Harvard University were asked whether it was possible to make a fully autonomous soft robot. In attempting to design one, they moved away from electronics, batteries, and motors to a structure that could be controlled by microfluidics.

The Octobot they produced provided the mechanics and core structure of an octopus-like robot made from a sandwich of materials that are not very different from the silicone caulk used to line bathroom sinks. However, the construction is much more complex. As an experimentalist with long-term involvement in three-dimensional (3D) printing, Lewis and her colleagues took advantage of the technology's ability to build complex structures in layers. As it forms each layer, the 3D printer works around the voids that will become microfluidic channels and pneumatic pipes used to fuel and distort the limb's shape.

Motive power for the Octobot's limbs comes from a supply of hydrogen peroxide in the robot's body. Microfluidic channels controlled by a network of tiny structures analogous

ACM News

Flexible Displays Enter the Picture

"Small, thin, and sleek" has long been the mantra for designing smartphones, smartwatches, tablets, and laptops. A focus on usability has driven advances in form factors. Yet, one thing has stayed the same: virtually all computing devices remain rigid.

It is a challenge that has vexed researchers struggling to develop electronic displays that can bend, fold, flex, and roll—while containing batteries, circuits, and other components. Ultimately, every advance has led to the same dead end: a display that cannot stand up to the rigors of everyday use.

However, that situation is about to change. After decades of research and false starts, manufacturers are introducing products with flexible displays.

Royole Corp. recently unveiled a smartphone with a flexible screen that allows the device to be folded like a billfold; the product has been available in China and

the U.S. since December 2018.

Meanwhile, Samsung plans to introduce a smartphone with a flexible display this year; others are incorporating flexible designs into products as well.

Says Vladimir Bulovic, a professor of electrical engineering at the Massachusetts Institute of Technology (MIT), "Flexible formats can be applied to many devices."

BEND, DON'T BREAK

New materials, better production methods, and other advances in technology have raised hopes of viable flexible products. The underlying OLED technology is now at a point where it works well, but encasing the displays in plastic or ultra-thin glass remains a challenge. Samsung's foldable smartphone, for example, features an interior screen that uses a composite polymer transparent material to encase a bendable AMOLED display.

The manufacturer claims the Samsung Infinity Flex Display can open and close 300,000 times without suffering damage.

Other companies are also moving flexible products from the research lab to production. Royole's FlexPai device features a 7.8-inch 1440p AMOLED display supported by a hinge that allows the device to flex to almost any desired angle. Royole also has partnered with Airbus to produce flexible electronics for aircraft, and plans to produce clothing with display technology.

Gregory Raupp, Foundation Professor of Chemical Engineering and founding director of the Flexible Display Center at Arizona State University, says flexible technology could impact fitness trackers, smart watches, Internet of Things devices, consumer electronics, and industrial control systems.

Yet, some questions remain. For one, "How do you deploy

that flimsy, plastic display into a product that will be robust and that the user won't damage by flexing it too much?" asks Raupp.

For another, "No one has attempted to produce flexible displays on a larger scale. Mass production creates distinct difficulties," says William Stofega, program director for Mobile Device Technology and Trends at IDC.

Finally, and perhaps most importantly, will the public desire flexible devices? "The social response to this technology is a complete unknown," Bulovic says.

Concludes Raupp, "We're now at a point where the manufacturing problems have largely been overcome and it is a question of innovating and integrating flexible displays in new, unique, and novel ways. It's up to the design community to transform ideas into reality."

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

to electronic logic gates to implement components such as oscillators convey the peroxide to reaction chambers. Each chamber contains a platinum catalyst, which splits the molecule into water and oxygen. The resulting gas drives pneumatic actuators that move the limbs, although they can do little more than twitch.

Soft robots may overcome the power problem by operating in environments where gravity is not as big a problem as it is on land. The microgravity of space may be one obvious habitat for them, but robots made from elastomers and pumps already find the going much easier under water.

Former Florida Atlantic University (FAU) student Jennifer Frame chose the jellyfish as the biological model for her thesis. Named JenniFish, the robot uses stubby plastic tentacles harnessed to hydraulic pumps powered by a battery to mimic the pulsing action of the invertebrate's body as it moves. Able to swim untethered in the ocean, it is flexible enough to squeeze through narrow orifices. In experiments, the robot would collide with the edges of a hole, then generate enough thrust to push its appendages to its side and squeeze through.

Another situation in which soft robots could perform well is in a different fluid environment: inside the human body. In work performed in Europe before moving to Stanford University in 2016, Stanford postdoctoral fellow Giada Gerboni (who works in surgical robotics) developed a soft endoscopic camera for surgery. Now Gerboni is focusing on soft robots that can enter the body, move around, and perform microsurgery operations without direct intervention from human surgeons. She describes it as a very flexible needle that can steer around parts of the body and into organs with minimal disruption.

In further work by the group led by FAU associate professor Erik Engeberg, the JenniFish has helped demonstrate how material choices form a key part of design for liquid environments. The algorithm needed to make the robot swim in different directions is relatively simple: it takes advantage of the way the polymer limbs are shaped to bend in on themselves when flexed. By changing the plastics used for the top and bottom surface of the JenniFish

Gerboni is focusing on soft robots that can enter the body, move around, and perform microsurgery operations without direct intervention from human surgeons.

tendrils, swapping hard for soft versions, the FAU team found there was an optimum combination for thrust: with both top and bottom being made of a material with floppiness similar to that of a mouse pad, being pulsed a little less than once a second.

In addition to the challenges robotists face with materials and power delivery, Josh Bongard, associate professor at the University of Vermont, says control presents further problems. "Exploiting the capabilities of soft robots is a very non-intuitive thing for human engineers to do. The mathematics that we've developed over decades for designing and controlling traditional robots made up of rigid links simply doesn't apply to [these] systems. In short: it's hard to design and control moving blobs."

In contrast to the inverse kinematics and closed-loop control that dominate fixed-function robots, Bongard proposes harnessing evolutionary programming coupled with machine learning to develop novel control methods for producing movement that take into account how plastic materials bend and compress under force.

In these simulations, the design starts with a basic shape made from blocks with different levels of stiffness and mobility. Evolutionary algorithms gradually change the properties of different blocks until the robot is able to move. The algorithms tune their response to the way materials flex under strain in different positions using machine-learning algorithms such as neural

networks. Sometimes, the simulated machines use the equivalent of body fat to help leverage the effects of whatever type of motion the robot adopts. Bongard says, "Our evolutionary algorithms often find very non-intuitive designs, such as one that has a hump on its back and uses it to throw its weight forward to make movement more efficient."

The work tends to agree with results such as those from FAU: softer robots perform better under water. Those simulated on land tended to require stiffer structures. Bongard now plans to take the work to physical robots in a collaboration with Kramer-Bottiglio and her team on a project backed by the National Science Foundation. The simulations will help create configurations for the robotic skins and the underlying shapes the skins are wrapped around.

Although soft robots have a long way to go before becoming autonomous enough to deliver on the promise of lighter, more functional machines, research and development is gradually bringing together the types of control and materials science that will make them work well on Earth and beyond. C

Further Reading

Rus, D., and Tolley, M.T.
Design, Fabrication and Control of Soft Robots
Nature 521, Number 7553, pp467-475 (2015)

Booth, J.W., Shah, D., Case, J.C., White, E.L., Yuen, M.C., Cyr-Choiniere, O., and Kramer-Bottiglio, R.
OmniSkins: Robotic Skins That Turn Inanimate Objects into Multifunctional Robots,
Science Robotics, Vol 3, Issue 22, eaat1853 (2018)

Frame, J., Lopez, N., Curet, O., and Engeberg, E.D.
Thrust Force Characterization of Free-Swimming Soft Robotic Jellyfish,
Bioinspiration & Biomimetics, Volume 13, Number 6 (2018)

Corucci, F., Cheney, N., Giorgio-Serchi, F., Bongard, J., and Laschi, C.
Evolving Soft Locomotion in Aquatic and Terrestrial Environments: Effects of Material Properties and Environmental Transitions,
Soft Robotics, 5(4): 475-495 (2018)

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

The Future of Data Storage

Research into next-generation storage techniques marches forward, yet tape remains the most viable, dependable medium.

OVER THE DECADES, computer storage has encompassed a variety of technologies, including punch cards, floppy disks, tape, hard drives, and flash technology. In every instance, the objective is the same: keep data accessible and available for the future. These advances in speed and capacity have helped today's sophisticated computing frameworks take shape. However, despite these gains, a simple but sobering fact emerges: "Tape remains a popular and preferred way to back up data," explains Robert Grass, a professor of chemistry at ETH Zurich (the Swiss Federal Institute of Technology) and a leading expert on nanotechnology.

Consider: When a software bug destroyed the email boxes of Gmail users in 2011, Google turned to tape to restore the data. The company spent more than 30 hours painstakingly recreating the accounts. Other companies and government organizations have encountered similar circumstances—and many continue to rely on tape. The reason? Tape remains inexpensive, the data on a tape remains accessible longer than on other media, and tape is remarkably easy to use and manage, while offering security benefits. "It is not an accident that tape remains in use," Grass says.

Although engineers continue to eke out further performance and capacity gains from hard drives and flash storage—and researchers are developing next-generation technologies such as DNA storage, crystal etching techniques, and molecular storage that could hold massive amounts of data on a small object for hundreds of thousands of years or longer—tape continues to march on, and on, and on. In 2017, for example, tape manufacturers shipped more than 1 million petabytes



of Linear-Tape-Open (LTO), a widely used magnetic tape data storage technology. This is about five times the volume of tape shipped in 2008.

Says Grass, "Tape may seem old-fashioned, and even obsolete, but from a purely economic point of view, it is the most cost-effective and efficient way to store data." As a result, "It isn't going to disappear anytime soon."

Recorded History

The evolution of tape has been nothing less than remarkable. In 1951, computing pioneers John Adam Presper "Pres" Eckert Jr. and John Mauchly introduced the world's first tape storage system: the UNISERVO tape drive for the UNIVAC computer. The device, which relied on 1/2-inch metal tape, was heavy and slow. It recorded data to eight channels at a density of 128 bits per inch. Moving at 100 inches per second, the tape delivered a practical transfer rate of about 7,200 characters per second. In contrast, today's tape devices transfer data at speeds

as high as 800 megabytes per second, while hard drives deliver a write speed of about 50 to 120 megabytes per second, and solid-state drives (SSDs) write data at rates of 200 to 550 megabytes per second.

By the 1970s, tape reels, cartridges, and cassettes had become the de facto way to back up and store data for both personal computers and enterprise systems. The limited capacity of floppy disk drives, and space and cost limitations imposed by early hard disk drives, kept tape at the forefront. Yet, even when disk technologies advanced exponentially and other storage technologies emerged, such as flash storage, the demand for tape didn't subside. "It has remained the standard for some very good reasons," says Reinhard Heckel, an assistant professor in the Department of Electrical and Computer Engineering at Rice University.

There are a number of technical and practical reasons tape has refused to fade into history. First and foremost,

the medium is considerably less expensive than disk or flash storage. Part of the reason is that, unlike disks, one tape machine can accommodate an unlimited number of tape drives or cartridges. An analysis conducted by BackupWorks.com indicates that equivalent levels of backup for tape versus disk results in about 4x cost savings for devices.

There are other cost benefits as well. These include a more than 2x savings in operational costs, and upward of 10x savings in total power and cooling costs. Today, a robotic tape library can contain upward of 278 petabytes of data; the same data stored on CDs would require almost 400 million discs.

Tape also delivers efficiency advantages. All devices, when they write bits to storage, produce an “unrecoverable bit error,” which occurs because the device writes a “1” instead of a “0” or vice versa. Error-correction methods do not make the problem completely go away. The result for a commonly used tape format such as an LT0-7/8 is a bit error rate of 1:1018, which is approximately one error for every 1.25 exabytes (EB). Other enterprise-class and consumer drives perform at an error rate of about 1:1016, which translates to an error every 125 terabytes (TB). Although error-correction codes for storage technologies have improved over the years, tape is about 100 times more accurate than the best hard drives, and about 10 times better than the latest solid-state devices (SSDs).

“The challenge with any storage technology is to reduce error codes,” Heckel says. In a practical sense, this means that tape systems are more dependable than other technology solutions. The greater the unrecoverable bit error rate, the greater the risk of loss of data, along with other errors and problems, including a system seeing two bad drives simultaneously.

Yet, there’s still another consideration. Modern tape cartridges fail at a rate about five orders of magnitude less frequently than hard drives—and tapes in storage require no moving or mechanical device.

Finally, tape offers the added appeal of creating an air-gapped environment when they are not in use. This

There are a number of technical and practical reasons why tape has refused to fade into history.

makes a tape library highly secure, as long as it is kept physically protected.

Beyond Tape

Tape is not ideal for every situation. Recovery from a tape backup can be slow and somewhat cumbersome. Finding specific files can prove vexing. If incorrectly stored, tapes can succumb to environmental damage or become demagnetized.

There’s also a bigger problem with all current storage technologies. A typical hard drive will operate only about three to five years before failing. Portable disk storage technologies such as CDs and DVDs generally hold data for 10 to 25 years, while flash storage—which includes drives, cards, and SSDs—degrades with use, rather than with age. This means that the more a user writes and rewrites to the device, essentially using it for its intended purpose, the greater the risk of failure. Future improvements in hard drives or flash technology are likely to produce only marginal performance gains, Heckel points out. However, tape, stored under ideal conditions, can last 30 to 50 years—and perhaps even longer. Although none of these technologies can compare to the lifespan of paper stored under ideal conditions (about 500 years), tape emerges as a clear winner.

Yet there’s still another long-term challenge: many existing storage technologies are butting up against physical and logical limits. It is increasingly difficult to add speed and capacity through more heads, platters, or microchips. A handful of technologies may help boost the power and scale of hard drives, for example. These include heat-assisted magnetic recording (HAMR) and microwave-assisted magnetic recording (MAMR). Both of

ACM Member News

PURSUING MULTI-PARTY COMPUTATION



Tal Rabin, a Distinguished Research Staff Member and manager of Cryptographic Research at IBM’s Thomas J. Watson Research Center in Yorktown, NY, was born in Massachusetts and grew up in Jerusalem, Israel.

Rabin earned bachelor of science, master of science, and doctorate of science degrees from the Hebrew University of Jerusalem, all in computer science. After obtaining her Ph.D., Rabin served as a postdoctoral fellow at the Massachusetts Institute of Technology for several years, before joining IBM’s T.J. Watson cryptography group.

Her research interests are in various areas of cryptography, with a primary focus on multi-party computation, a way for two or more parties to together compute a function over their inputs, while keeping the individual inputs private.

Rabin is currently engaged in designing a distributed protocol that would enable people to submit a sexual harassment complaint, which would then identify other potential victims of that specific harasser. Rabin explains that victims might be more likely to move forward with such a complaint if they knew a group of people would also be involved, so they would not have to proceed alone.

“This is the beginning of the road for having actual applied multiparty computation algorithms,” Rabin says, adding that the next few years will be exciting as these applications come to fruition.

Rabin is also interested in helping women advance in computer science by getting more women involved at the undergraduate and graduate levels, as well as in the workplace. She acknowledges that she does not have the answers as to how this should be done, but she hopes it will be a community effort.

—John Delaney

these methods allow smaller regions of a disk to be magnetized, resulting in higher capacity. However, these approaches also boost costs, and result in density scaling gains of only about 15%.

Researchers are now exploring other technologies that could one day replace disks, tapes, and flash memory—or at least supplement them for specific uses. Underpinning this is the fact that up to 90% of the data generated by computers and other digital systems is never accessed again; it simply lies idle, consuming ever-growing mountains of storage media or servers.

Likewise, there is the issue of hard drive capacity. “The problem with today’s systems is that they deliver no more than one terabyte per square inch,” says Karthik V. Raman, a former IBM research scientist who now leads a research team at the Tata Institute of Fundamental Research (TIFR) in India.

The net effect is that current storage technologies—particularly disk-based servers and systems—consume massive amounts of physical space, particularly when they involve large numbers of devices and media, such as tapes or disks. Yet, even tape produces enormous volume of physical objects. It is estimated that humans produce approximately 2.5 quintillion bytes of data each day and, overall, nearly 3 zettabytes of data exist in the digital world. All these bytes require increasingly large data centers that consume massive amounts of energy, along with other resources.

However, Raman points out, “Creating new types of storage with greater capacity doesn’t solve the problem by itself. There’s a need to develop better ways to direct and redirect data for faster processing.” Heckel and Grass, for instance, have focused on using DNA as a data storage mechanism. The idea, first presented by George Church, a molecular biologist and geneticist at Harvard Medical School, involves writing data to DNA material, which could conceivably store that data for hundreds of thousands, or even millions, of years. Church says the purpose of DNA storage “isn’t to reinvent the hard drive, it’s to introduce a medium that is ideal for archiving and long-term storage.”

Others are taking a different tack for keeping data intact for long pe-

Researchers are now exploring other technologies that could one day replace disks, tapes, and flash memory—or at least supplement them for specific uses.

riods of time. For example, Peter G. Kazansky, a professor at the University of Southampton in the U.K., has developed a method that uses an ultrafast short-pulse laser to etch data into the bulk of silica material. “A single disk using this technology could store 360 terabytes of data, compared to a Blu-ray disk that can store about 45 gigabytes,” he says. Moreover, the data would potentially stay on the disk for approximately 14 billion years. The project has caught the eye of Microsoft, which is working to produce a commercially viable version of the technology within a decade (and its Project Silica is focusing on ways to use the technology in the cloud).

Tape Prevails

Remarkably, all storage devices and use cases eventually lead back to tape—at least, for the foreseeable future. While tape is not as flexible or convenient as hard drives, SSDs, and other media, it remains cost-effective and highly reliable. What’s more, advancements in tape continue to outpace other storage technologies. In 2017, IBM and Sony announced a new magnetic tape system capable of storing 201 gigabytes of data per square inch in a single palm-sized cartridge. The technology has a theoretical limit of 330 terabytes per square inch. The world’s largest hard drives, on the other hand, require twice the physical space, but hold only 12 terabytes per square inch. The most advanced

SSDs hold about 60 terabytes per square inch.

Many experts say the practical and cost advantages tape has over hard drives and other storage technologies will likely grow over the next several years. Tape won’t ever threaten hard drives and SSD for dominance, but it will remain at the center of storage—and provide a strong insurance policy for the likes of Google. Mark Lantz, manager of Advanced Tape Technologies at IBM Research Zurich, noted in an August 2018 *IEEE Spectrum* article that researchers continue to boost the density and capacity of tape, and the trend will continue for some time. “Tape may be one of the last information technologies to follow a Moore’s Law,” he wrote.

To be sure, tape remains viable and valuable—and the situation is not likely to change anytime soon. Concludes Grass, “Other emerging technologies will eventually change the way data is stored. But for archival data storage, tape is the technology to beat.”

Further Reading

Blawat, M., Gaedke, K., Hütter, I., Chen, X., Turczyk, B., Inverso, S., Pruitt, B.W., and Church, G.M.

Forward Error Correction for DNA Data Storage. Procedia Computer Science, Volume 80, 2016, pp. 1011-1022.

<https://www.sciencedirect.com/science/article/pii/S1877050916308742?via%3Dihub>

Zhang, J., Čerkauskaitė, A., Drevinskas, R., Patel, A., Beresna, M., and Kazansky, P.G.

Current Trends in Multi-Dimensional Optical Data Storage Technology, Asia Communications and Photonics Conference 2016, Current Trends in Multi-Dimensional Optical Data Storage Technology, Wuhan China, November 2-5, 2016.

<https://doi.org/10.1364/ACPC.2016.AF1J.4>

Zhang, J., Čerkauskaitė, A., Drevinskas, R., Patel, A., Beresna, M., and Kazansky, P.G.

Eternal 5D Data Storage by Ultrafast Laser Writing in Glass. Proc. SPIE 9736, Laser-based Micro- and Nanoprocessing X, 97360U (4 March 2016); doi: 10.1117/12.2220600; https://doi.org/10.1117/12.2220600

Heckel, R., Mikutis, G., and Grass, R.N.

A Characterization of the DNA Data Storage Channel, eprint arXiv:1803.03322. March 2018. https://arxiv.org/abs/1803.03322v1

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

The Fine Line Between Coercion and Care

Employers monitoring their workforce must balance productivity and security considerations with employee privacy and welfare concerns.

EMLOYEE MONITORING IS an age-old practice in industrial society, harking back to manual timesheets. It has since developed in line with technology breakthroughs into a highly sophisticated process that is actively practiced by employers, and not always welcomed by employees.

Multinational professional services firm Deloitte notes the prevalence of employee monitoring in a 2018 Global Human Capital Trends article ("People data: How far is too far?") authored by the consultancy's human capital leaders. The article states, "Use of workforce data to analyze, predict, and help improve performance has exploded over the last few years. But as organizations start to use people data in earnest, new risks as well as opportunities are taking shape." The opportunities include collecting employee data to address issues such as productivity and employee engagement to make better business decisions. The risks include breaching data privacy and alienating employees.

Michel Anteby, associate professor of organizational behavior and sociology at Boston University's Questrom School of Business, relates a rather chilling tale of employee monitoring. In a study he started in 2011 and reported in 2018, he tracked the experiences and resistance strategies of security screening personnel subject to CCTV surveillance at a large urban airport. Ironically, the U.S. Transportation Security Administration (TSA) that employed the personnel and set up the surveillance to strengthen managerial control of employee theft and reassure the vast majority of employees that they were not responsible for theft, ended up with a disillusioned workforce that felt a sense



of visibility of behavior, but a lack of management notice as individuals. This led the staff to engage in invisibility practices in an attempt to go unseen and remain unnoticed.

Explains Anteby, "Increasing surveillance made the employees feel watched in a coercive way, caught out if something went wrong, but not rewarded when things went right. So they adopted invisibility practices to stay out of sight of managers watching them on CCTV."

The employees took two approaches. First, they practiced invisibility of behavior, which allowed them to temporarily escape the scrutiny of management by taking extended breaks, staying out of sight as much as pos-

sible physically, and when asked to move to a different gate at the airport, taking a route that took twice as much time to complete as necessary.

The second approach, which Anteby describes as "invisibility of self," involves employees making efforts to "disappear in plain sight." He explains, "Even if management could see the employees, they moved about under the radar, showing no emotion, nothing identifiable about their clothes or demeanor, nothing that marked them out as individuals. There were about 1,000 employees, and they were very good at this."

Here the tale takes another ironic twist, with TSA management saying because it could not see enough of the

employees, it was implementing more CCTV cameras. Anteby's study, which was funded by Boston University in collaboration with airport security staff, ended at about this time, but he suggests the employees had developed a skillset that would continue to keep them under the radar. Anteby concluded, "The airport management was surprised by our findings, it was trying to improve the workplace."

Coercive surveillance can also be seen in police departments, where individual officers feel surveillance by senior officers is about identifying misbehavior, rather than developing better officers. These types of surveillance can be clearly described, but the reality is that technological development has made surveillance prevalent in the workplace, yet sometimes covert. There are plenty of surveillance software vendors in the market that can provide the technology to monitor employee activity and communication, including Internet and app usage, email, messaging, computer screen recording, capturing key strokes (or the lack of them), telephone use, video and audio, location tracking with access cards, and vehicle tracking using the global positioning system (GPS).

For employers, surveillance can help firms with everything from the elimination of time-wasting practices and the setting of salaries appropriately to the protection of intellectual property and ensuring the company has the right tools and equipment in place to optimize employee performance and productivity. Surveillance can also provide information about specific employees, which is where the point of contention often lies.

With so much technology installed in business and industry, employees may not be aware that they are being tracked, perhaps through the use of electronic keycards to access elevators or offices on a particular floor, or a closed-circuit video system mounted in an office garage to identify the license plates of employee vehicles to allow for secure entry and exit. While these devices are usually installed for security purposes, they can be used equally well to track when staff arrive at, and leave, the office—covert surveillance, even if it not intended

as such, that will only come to light when management questions employee behavior.

Even if employees are aware they are being monitored, surveillance can damage morale, cause stress, and raise questions about privacy. According to Kate Bischoff, an employment attorney and human resources consultant as well as a senior certified professional of the Society for Human Resource Management (SHRM), the consequence of lower morale, higher stress, and less privacy is poor retention. Higher turnover is also bad for morale, thus amplifying the issue, as well as being costly and time-consuming for employers, who then must go through the laborious processes of hiring and training new employees.

So, where can the line be drawn between what Anteby calls coercive and caring surveillance? How can employees push back against employer surveillance schemes, and what are the legal parameters? These questions do not have precise, single answers and depend on circumstances.

Anteby suggests the line between coercive and caring surveillance does not depend on types of jobs, but rather on the way surveillance is carried out. He describes coercive systems as those that seek to catch employees when they don't follow the rules. Caring systems he describes as those that provide developmental coaching to employees on how they can improve their skills.

If an employee contract includes acceptance of a certain level of sur-

veillance, a fairly clear line can be drawn; otherwise, when employer surveillance becomes employee abuse is a moving target. While unionized employees have more push-back than others against employers with the potential to strike, this will always lead to uncertain, and sometimes detrimental, outcomes.

The law provides little more certainty. In the U.S., federal law generally gives employers the right to monitor employees as they perform their work, but probably not to monitor them in private areas such as bathrooms and locker rooms, while state law takes a more consensual approach. Similarly, the U.K. offers guidelines rather than prescriptive measures. Minal Backhouse, managing director of Backhouse Solicitors, an Essex, U.K.-based specialist in employee law, says, "Monitoring employees is an accepted and useful tool in a well-run business. But to stay on the right side of the law, proper documentation should explain what is being monitored and why, and trust should be maintained through clear communication with employees."

Albert Gidari, director of privacy at the Center for Internet and Society at Stanford University Law School, describes today's levels of surveillance as "a technology story of normalization of invasiveness into people's private lives," and argues that there is a trade-off between surveillance and life improvements such as improved communication and sharing of information.

Noting the use of the instant messaging service Slack, he says that in the U.S., the use of Slack is not seen as an invasion of privacy, but instead as a workplace enhancement. Anteby also mentions the use of shared workplace tools such as Slack, but notes that employees, particularly senior employees, often communicate outside Slack using text or phone to avoid some of their messages being open to monitoring by other staff.

Outside the office, tracking and surveillance have a long history in the transport and trucking industry. The UPS parcel delivery service, by way of example, uses computer analytics to monitor its delivery drivers and provide guidance on how to avoid wasting time and fuel on their routes.

Even if employees are aware they are being monitored, surveillance can damage morale, cause stress, and raise questions about privacy.

Truck drivers are among the most accepting subjects of monitoring, and often see surveillance as care, rather than coercion. Steve Viscelli, senior fellow at the Kleinman Center for Energy Policy, and a lecturer in the Department of Sociology at University of Pennsylvania, chronicles decades of technology approaches to monitoring in the trucking industry, noting early disk recording systems that could be used after the fact to observe basic information such as driving time and how fast assignments were completed. They could not be used for direct monitoring, making it impossible to decipher when a truck would turn up at a location and plan the driver's next journey. In the late 1980s, satellite solutions emerged that could monitor trucks in real time and allow future work to be scheduled.

With many drivers joining the trucking industry as a means of gaining independence, getting away from the boss in the office, and being paid for the amount of work they do, Viscelli expected satellite solutions to have a profound effect on drivers' desire for autonomy. However, he says: "I was wrong. Most truckers didn't mind. For conscientious workers, monitoring backed up what they were doing and could show if they had a problem." Other technologies packed into solutions provided instructions, perhaps what drivers had to pick up next, enhancing the earnings of drivers that are paid by the mile.

By the late 1990s, these systems, also known as Qualcomms (as they were based on Qualcomm chips), dominated big fleets. However, they could be 'gamed' to underreport driving hours in electronic logs required by the U.S. Department of Transportation, a situation that was not tenable as fatigued drivers caused more accidents.

Viscelli describes GPS as the 'new frontier' of monitoring, as it tracks trucks when they are moving and when they are at a location, helping both employers and employees reduce downtime. Forward-facing video is also finding favor with drivers as it helps to reconstruct accidents and can show plainly that a car, rather than a truck, was at fault, which is most frequently the case,

Truck drivers are among the most accepting subjects of monitoring, and often see surveillance as care, rather than coercion.

says Viscelli. Less popular are driver-facing videos, and most recently, computer programs that use videos to watch drivers' eyelids to see if they are micro-sleeping or fatigued. That said, Viscelli says, "Experienced drivers already take monitoring in their stride, and it will become ubiquitous once safety benefits are clear and documented."

While questions continue to be raised about the rights and wrongs of employee monitoring, it appears likely they will be swept aside by employers with increasing technological ability to gather more information, leaving employees with little choice but to accept, sometimes reluctantly, that workplace monitoring is all part of the deal. □

Further Reading

People data: How far is too far?, Deloitte Insights 2018 Global Human Capital Trends, <https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2018/people-data-analytics-risks-opportunities.html>

Antebi, M., and Chan, C.K. A Self-Fulfilling Cycle of Coercive Surveillance: Workers' Invisibility Practices and Managerial Justification, Organization Science, Vol. 29, No. 2, 2018 <https://doi.org/10.1287/orsc.2017.1175>

Viscelli, S. The Big Rig: Trucking and the Decline of the American Dream, University of California Press, 2016 <https://www.steveviscelli.com/book>

Sarah Underwood is a technology writer based in west London, U.K.

© 2019 ACM 0001-0782/19/4 \$15.00

Milestones

Newest NAE Members Include Computer Scientists

Among the 86 new members and 18 foreign members recently elected to the National Academy of Engineering (NAE) are five computer scientists.

Election to the National Academy of Engineering is among the highest professional distinctions accorded to an engineer.

Individuals in the newly elected class will be formally inducted during a ceremony at the NAE's annual meeting in Washington, D.C., on Oct. 6, 2019.

The computer scientists newly elected to the National Academy of Engineering (and the reason for their election) are:

Joseph Y. Halpern, Joseph C. Ford Professor of Engineering, computer science department, Cornell University, Ithaca, NY, for methods of reasoning about knowledge, belief, and uncertainty and their applications to distributed computing and multiagent systems.

Monica S. Lam, professor, computer science department, Stanford University, Stanford, CA, for contributions to the design of advanced compiler and analysis systems for high-performance computers.

Robert T. Morris, professor, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, for contributions to programmable network routers, wireless mesh networks, and networked computer systems.

Margo I. Seltzer, Herchel Smith Professor of Computer Science, School of Engineering and Applied Science, Harvard University, Cambridge, MA, for engineering contributions to databases, file systems, and operating systems.

Claire J. Tomlin, Charles A. Desoer Chair and professor, electrical engineering and computer sciences, University of California, Berkeley, for contributions to design tools for safety-focused control of cyberphysical systems.

DOI:10.1145/3313138

Mari Sako

Technology Strategy and Management

Free Trade in a Digital World

Considering the possible implications for free trade, traditionally based on non-digital goods, for a modern global economy that is increasingly based on intangible products and services enabled by digital technologies.

ASPECTER IS HAUNTING the globe—a specter of neo-nationalism and protectionism. But not all the global powers are united in an alliance to exorcise this specter, because its lead conjuror is the U.S.—the largest economy in the world. U.S. protectionism was presaged by an ironic juxtaposition at the World Economic Forum in January 2017, when China's President Xi Jinping championed the cause of a liberal economic order in contrast to President Trump's America First stance. Why has the world moved toward protectionism, and what is its impact on businesses and consumers? And how damaging is this phenomenon to our prosperity? This column considers what free trade has meant, and the impact of its demise.

This topic is particularly important to people involved in computing in an increasingly digital world. The history of free trade centers around non-dig-

ital goods and services, given the relatively late arrival of digital and digitally traded goods and services. That history helps explain the issue and draw implications for the digital world.

History of International Trade

Trading has a very long history underpinning ancient civilizations that thrived in Egypt, Greece, and Rome. Later, merchants in Venice and elsewhere traded by sea, but also along the Silk Road connecting China to the Middle East and the West. It was not until the 18th century, however, that the economists Adam Smith and David Ricardo articulated the modern notion of freedom of commerce and freedom of the seas. By then, the Dutch and the British, among other sea powers, cultivated colonies, and fueled a fierce debate between mercantilists and free traders, a debate that lasted well into the late 19th century. Mercantilists believed strong nation-states had to be self-sufficient

with secure local resources, trade surplus, and the accumulation of precious metals, particularly gold and silver in the form of bullion. Consequently, they regarded international trade as a zero-sum game.

Against this historical backdrop, the free trade system held together by multiple sovereign nation-states, as we know it today, is a 20th-century creation. The Great Depression in the 1930s led to protectionist moves by major industrial nations. But thereafter, the General Agreement on Tariffs and Trade (GATT) since 1948 and the World Trade Organization (WTO), which superseded GATT in 1995, facilitated significant reductions in quotas and tariffs. Consequently, international trade in merchandise surged to 17 trillion USD in 2017. Moreover, much of this volume is accounted for by trade in intermediate goods. This is a manifestation of the rise of global supply chains in a variety of assembly-based manufacturing



sectors including clothing, footwear, and electronics.⁴ In electronics, for example, components that are assembled into mobile phones zap across national borders more frequently than completely assembled mobile phones themselves. In the golden age of global supply chains, we have taken dispersed production locations for granted.

Free trade improves productive efficiency and increases consumer choice. The economic theory of comparative advantage also provides an indisputable logic behind why two nations that trade are both better off than if they do not trade. Nevertheless, free trade remains controversial because gains from trade are not evenly distributed within countries. Liberalizing international trade necessarily produces losers among workers in sectors facing import competition as they face job losses and declining wages. The resulting greater inequality begets a sense of injustice, and has led to a populist backlash in many countries. We need robust institutions to deal with redistributive consequen-

ces of free trade, and not all nation-states are up to carrying out that task.³

The Current Phase of Protectionism
Many commentators draw parallels between the post-2008 situation that we are currently experiencing and the 1930s. Just like in 1929, the global economy suffered a great financial crisis in 2008. Just like in the 1930s, the U.S. has imposed tariffs on imports in 2018. But so far, U.S. protectionism under Presi-

dent Trump has been unilateral, with only China retaliating. So unlike in the 1930s, no major industrial nation other than China has retaliated. Optimism is not warranted, however. Other countries have resorted to trade-distorting policies such as subsidies and local content requirements well before President Trump came onto the scene. In fact, the G20 nations—including China and the U.S.—first met in 2008 to forswear protectionism, but have been steadily increasing their policy interventions to tilt the commercial playing field in favor of domestic interests.²

The U.S. trade policy has been quite explicit about protecting U.S. jobs. In the 1980s, U.S. administrations focused on protecting jobs in the automobile and electronics sectors facing competition from Japan. In the 2010s, creating manufacturing jobs, of concern under the Obama administration, took a protectionist turn under Trump, starting with tariffs on steel and aluminum to protect steel and aluminum jobs. On the other side of the Atlantic,

Why has the world moved toward protectionism, and what is its impact on businesses and consumers?

ACM Transactions on Spatial Algorithms and Systems

ACM TSAS is a new scholarly journal that publishes high-quality papers on all aspects of spatial algorithms and systems and closely related disciplines. It has a multi-disciplinary perspective spanning a large number of areas where spatial data is manipulated or visualized.

The journal is committed to the timely dissemination of research results in the area of spatial algorithms and systems.



For further information or to submit your manuscript, visit tsas.acm.org

Britain's decision to exit the European Union—so-called Brexit—creates uncertainty. While Britain may be doing its best to keep its free trade flag flying, doing so at the same time as exiting a free trade zone undermines credibility.

From the perspective of developing countries, exporting raw materials and manufactured goods are as important as ever for their prosperity. In a neo-liberal world, the 19th-century German economist Frederich List's memorable phrase 'kicking away the ladder' resonated with many observers. In the past, rich industrialized nations got richer by using high tariffs before lowering them. How can we insist that poorer nations attempt to industrialize without such protective ladders to climb up?¹ In the 2010s, however, no nation is kicking away the ladder. Rather, import substitution and industrial policies—once seen to be the preserve of developing economies—are applied in different shapes and forms around the world. Moreover, developing nations continue to face significant barriers to trade in sectors other than manufacturing, notably exporting agricultural produce to developed nations. Thus, advocates of free trade have always tempered their demand by accommodating the need to secure national sovereignty and national interest.

The Digital Dimensions of Protectionism

The preceding analysis may give the impression the pendulum swinging between free trade and protectionism is all about geopolitics and conflict between nations. But that is only partially correct. Technology—particularly digital technology—has played a great part in enhancing international trade. The digital dimension is highly relevant to people who work in the computing world. In some cases our attention to this may be restricted to what might be called "pure" digital services or products. But often it involves services or products that depend on digital technology in their creation, delivery, or use. In fact, the rapid incorporation of the digital into many services and products could be of greater significance than anything "pure."

As noted here, GATT and WTO facilitated the unprecedented growth in international trade in the second half of the 20th century. The key enabling

technology for making free trade thrive was initially container shipping and intermodal transport, lowering the cost of physical transportation. Later, computing technology ushered in an era of call centers, offshore software development, and business service delivery from remote locations such as India. The technologies at the forefront—Computer Aided Design (CAD), Computer Aided Manufacturing (CAM), and the Internet—facilitated the dispersion of productive locations to design, manufacture, and provide after-sales service of complex products such as aircraft.

The power of new technologies to continue to promote geographically dispersed production networks is here to stay. But interestingly, new technologies may just as well be put to use to facilitate proximate design and production. When speed-to-market and reacting to consumer feedback in real time are important, locating design and production facilities close to final markets makes sense. There are signs of this happening. For example, the sportswear maker Under Armour has designers and manufacturers located under one roof in Baltimore, MD, USA. Adidas has one Speedfactory using robots in Ansbach, Germany, and another similar factory in Atlanta, GA, USA; and Nike's New Manufacturing partnership with Flextronics located in U.S. states California and Tennessee appears to be more and more about design and innovation, and less about low cost. And of course, these proximate locations create onshore jobs, which happens to be compatible with the climate of neo-nationalism and protectionism.

It is the nature of the technology in 3D printing, robots, flexible manufac-

The power of new technologies to continue to promote geographically dispersed production networks is here to stay.

turing systems, and AI, that makes it possible to design and produce close to end users. First, economies of scale are less important, enabling extreme customization possible to suit consumer demand. Second, automation and machine learning, incorporating elements of AI, diminish the advantage of locations with low labor costs. And third, as noted previously, the technology facilitates the geopolitical climate of neo-nationalism and protectionism. The digital technology might be neutral to the necessity of distant or proximate production and trading, but the nature of trade can nevertheless be a significant matter.

Future of Free Trade

What does the analysis presented in this column lead us to think about the future of free trade? Given the potential of digital technology, one may remain optimistic. This optimism does not rest on digital technology's capacity to override the shortcomings of politics; it comes from the potential to use digital technology to facilitate shifts in politics and consumer preferences.

Pessimists may focus on a possible future with more protectionist retaliations by other nations, just like in the 1930s. More tariffs on semiconductor and telecom equipment, for instance, would mean rising costs of producing ICT equipment, and a decline in global supply chains in that sector. Moreover, even as more and more of the economy becomes digital and less and less about physical trading of goods, 'virtual' will remain connected to physical locations. Amazon's acquisition of Whole Foods in 2017 attests to the importance of prime urban locations as delivery nodes. Even in a pure digital sphere, governments regulate data flows across national borders due to concerns over privacy and cyber security. So, free trade in digital products and services may be the ideal, but national governments remain vigilant about where and how data is held.

Some optimists may think that in the struggle between the physical world (with geopolitics pushing governments away from openness) and the digital world (fostering greater social integration and connectivity), the digital would win out in the end. However, this rests on a somewhat false di-

Significant segments of the global economy should lament the end of the golden era of global supply chains.

chotomy, particularly if we focus on the producer side to meet demand. In the past, the digital technology revolution was about enabling geographically dispersed production networks. In the future, digital technology can be about enabling proximity to customers. Agility and local relevance stem from using digital technology to realize greater interaction, high customization, and resource saving.

Conclusion

Free trade as an ideal had important application to manufactured goods in the second half of the 20th century. Its incorporation into services is less clear. It may not spread beyond our current stage to other sectors, including intangibles such as services and products enabled by digital technologies. Significant segments of the global economy should lament the end of the golden era of global supply chains. The trigger for the beginning of the end may be the new enabling technologies of 3D printing, robots, and AI, as much as the geopolitics of today. Neo-nationalism and protectionism need not be the path of the future, and digital technologies may well transform the direction taken by world trade. □

References

- Chang, H.-J. *Kicking Away the Ladder: Development Strategy in Historical Perspective*. Anthem Press, London, 2002.
- Evenett, S.J. and Fritz, J. Brazen unilateralism: The U.S.-China tariff war in perspective. *The 23rd Global Trade Alert Report*. CEPR Press, London, U.K., 2018; <http://www.globaltradealert.org>
- Rodrik, D. *Straight Talk on Trade*. Princeton University Press, Princeton, NJ, USA, and Oxford, U.K., 2018.
- Sako, M. Driving power in global supply chains. *Commun. ACM* 54, 7 (July 2011), 23–25.

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

Copyright held by author.

Calendar of Events

April 3–4

SOSR '19: Symposium on SDN Research, San Jose, CA, Sponsored: ACM/SIG, Contact: Eric Rozner, Email: erozner@gmail.com

April 8–12

SAC 2019: The 34th ACM/SIGAPP Symposium on Applied Computing, Sponsored: ACM/SIG, Limassol, Cyprus, Contact: Chih-Cheng Hung, Email: chung1@kennesaw.edu

April 13–14

VEE '19: 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments, Providence, RI, Co-Sponsored: ACM/SIG, Contact: Jennifer B. Sartor, Email: jbsartor@gmail.com

April 13–17

ASPLOS '19: Architectural Support for Programming Languages and Operating Systems, Providence, RI, Co-Sponsored: ACM/SIG, Contact: Iris Bahar, Email: iris_bahar@brown.edu

Apr 14–17

ISPD '19: International Symposium on Physical Design, San Francisco, CA, Sponsored: ACM/SIG, Contact: Ismail Bustany, Email: ismail.bustany@gmail.com

April 15–18

CPS-IoT Week'19: Cyber-Physical Systems and Internet of Things Week 2019, Montreal, QC, Canada, Sponsored: ACM/SIG, Contact: Xue Liu, Email: xueliu@cs.mcgill.ca



DOI:10.1145/3313140

George V. Neville-Neil

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

Kode Vicious

Know Your Algorithms

Stop using hardware to solve software problems.

Dear KV,

My team and I are selecting a new server platform for our project and trying to decide whether we need more cores or higher-frequency CPUs, which seems to be the main trade-off to make on current server systems. Our system is deployed on the highest-end systems and, therefore, the highest-frequency systems we could buy two years ago. We run these systems at 100% CPU utilization at all times. Our deployment does not consume a lot of memory, just a lot of CPU cycles, and so we are again leaning toward buying the latest, top-of-the-line servers from our vendor. We have looked at refactoring some of our software, but from a cost perspective, expensive servers are cheaper than expensive programmer time, which is being used to add new features, rather than reworking old code. In your opinion, what is more important in modern systems: frequency or core count?

Richly Served

Dear Served,

I really wish I knew who your vendor is, so I could get a cut of this incredibly lucrative pie. As the highest-end servers currently enjoy a massive markup, your salesperson probably appreciates every time you call.

The short answer to your question about frequency vs. core count



is, "You tell me." It seems as if you have spent little or no time trying to understand your own workload and have simply fallen for the modern fallacy of "newer will make it better." Even apart from the end of frequency scaling, it has rarely been the case that just adding more oomph to a system is the best way to improve performance. The true keys to improving performance are measurement and an understanding of algorithms.

Knowing your CPU is in use 100% of the time does not tell you much

about the overall system other than it is busy, but busy with what? Maybe it is sitting in a tight loop, or someone added a bunch of delay loops during testing that are no longer necessary. Until you profile your system, you have no idea why the CPU is busy. All systems provide some form of profiling so you can track down where the bottlenecks are, and it is your responsibility to apply these tools before you spend money on brand-new hardware—especially given how wacky new hardware has been in the

past five years, particularly as a result of NUMA (non-uniform memory access) and all the convoluted security mitigations that have sapped the life out of modern systems to deal with Spectre and the like. There are days when KV longs for the simplicity of a slow, eight-bit microprocessor, something one could understand by looking at the stream of assembly flying by. But those days are over, and, honestly, no one wants to look at cat videos on a Commodore 64, so it is just not a workable solution for the modern Internet.

Since I have discussed measurement before, let's discuss now the importance of algorithms. Algorithms are at the heart of what we as software engineers do, even though this fact is now more often hidden from us by libraries and well-traveled APIs. The theory, it seems, is that hiding algorithmic complexity from programmers can make them more productive. If I can stack boxes on top of boxes—like little Lego bricks—to get my job done, then I do not need to understand what is inside the boxes, only how to hook them together. The box-stacking model breaks down when one or more of the boxes turns out to be your bottleneck. Then you will have to open the box and understand what is inside, which, hopefully, does not look like poisonous black goo.

A nuanced understanding of algorithms takes many years, but there are good references, such as Donald Knuth's book series, *The Art of Computer Programming*, which can help you along the way. The simplest way to start thinking about your algorithm is the number of operations required per unit of input. In undergraduate computer science, this is often taught by comparing searching and sorting algorithms. Imagine that you want to find a piece of data in an array. You know the value you want to find but not where to find it. A naive first approach would be to start from element 0 and then compare your target value to each of the elements in turn. In the best case, your target value is present at element 0, in which case you have executed a very small number, perhaps only one or two instructions.

Personally, I never bother with the best case, because I always expect that, on average, everything will be worst case.

The worst-case scenario is that your target element does not exist at all in the array and you will have executed many instructions—one comparison for every element of the array—only to find that the answer to your search is empty. This is called a linear search.

For many data structures and algorithms, we want to know the best, worst, and average number of operations it takes to achieve our goal. For searching an array, best is 1, worst is N (the size of the array), and average is somewhere in the middle. If the data you are storing and searching is very small—a few kilobytes—then an array is likely your best choice. This is because even the worst-case search time is only a few thousand operations, and on any modern processor, that is not going to take a long time. Also, arrays are very simple to work with and understand. It is only when the size of the dataset grows into megabytes or larger that it makes sense to pick an algorithm that, while it might be more complex, is able to provide a better average number of operations.

One example might be to pick a hash table that has an average search time of one operation and a worst search time of N —again the number of elements in the table. Hash tables are more complex to implement than arrays, but that complexity may be worth the shorter search time if, indeed, searching is what your system does most often. There are data structures and search algorithms that have

been developed over the past 30 years with varying performance characteristics, and the list is too long, tedious, and boring to address in depth here. The main considerations are how long does it take, in the best, worst, and average cases, to:

- ▶ Add an element to the data structure (insertion time).
- ▶ Remove an element.
- ▶ Find an element.

Personally, I never bother with the best case, because I always expect that, on average, everything will be worst case. If you are lucky, there is already a good implementation of the data structure and algorithm you need in a different box from the one you are using now, and instead of having to open the box and see the goo, you can choose the better box and move on to the next bottleneck. No matter what you are doing when optimizing code, better choice of algorithms nearly always trumps higher frequency or core count.

In the end, it comes back to measurement driving algorithm selection, followed by more measurement, followed by more refinement. Or you can just open your wallet and keep paying for supposedly faster hardware that never delivers what you think you paid for. If you go the latter route, please contact KV so we can set up a vendor relationship.

KV

Related articles on queue.acm.org

KV the Loudmouth

<https://queue.acm.org/detail.cfm?id=1255426>

10 Optimizations on Linear Search

Thomas A. Limoncelli

<https://queue.acm.org/detail.cfm?id=2984631>

Computing without Processors

Satnam Singh

<https://queue.acm.org/detail.cfm?id=2000516>

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

Viewpoint

The Web Is Missing an Essential Part of Infrastructure: An Open Web Index

A proposal for building an index of the Web that separates the infrastructure part of the search engine—the index—from the services part that will form the basis for myriad search engines and other services utilizing Web data on top of a public infrastructure open to everyone.

THE WEB AS it currently exists would not be possible without search engines. They are an integral part of the Web and can also be seen as a part of the Web's infrastructure. Google alone now serves over two trillion search queries per year.¹¹ While there seems to be a multitude of search engines on the market, there are only a few relevant search engines in terms of them having their own index (the database of Web pages underlying a search engine). Other search engines pull results from one of these search engines (for example, Yahoo pulls results from Bing), and should therefore not be considered search engines in the true sense of the word. Globally, the major search engines with their own indexes are Google, Bing, Yandex, and Baidu. Other independent search engines may have their own indexes, but not to the extent that their size makes them competitive in the global search engine market.

While the search engine market in the U.S. is split between Google and Bing (and its partner Yahoo) with approximately two-thirds to one-third, respectively,¹⁰ in most European coun-



tries, Google accounts for more than 90% of the market share. As this situation has been stable over at least the last 10 years, there have been discussions about how much power Google has over what users get to see from the Web, as well as about anticompetitive business practices, most notably in the context of the European Commission's competitive investigation into the search giant.³

Search Engine Bias?

From the users' point of view, search engines are reliable and trustworthy sources, providing fair and unbiased results.⁸ However, it has been found that search results simply should not be considered "neutral." Some scholars argue that an unbiased search engine is simply not possible, as there is no ideal result set against which a bias can be measured.^{5,6} Therefore, I argue

that every search engine presents its own algorithmically generated view of the Web's content. Every such view can be different, and none of them are the definitive or correct one.

Problems that may arise from search engines' interpreting the world in certain ways include: reinforcing stereotypes, for example, toward women;⁷ influencing public opinion in the context of political elections (see, for example, Epstein and Robertson²); and preferring dramatic interpretations of rather harmless health-related symptoms.¹³

It seems, therefore, unreasonable to have only one (or a few) dominant search engines imposing their view on the Web's content, which is, on closer inspection, really only one of many possible views. Therefore, I argue for building an index of the Web that will form the basis for a multitude of search engines and other services that are based on Web data.

Three Major Problems

There are three major problems resulting from a search engine market where only a few competitors are equipped with their own index of Web pages:

- A search engine provides only one of many possible algorithmic interpretations of the Web's content. At least for informational queries (see Broder¹), there is no correct set of results, let alone one single correct result. For these queries, we usually find a multitude of results of comparable quality. While a search engine's ranking might provide some relevant results on the highest positions, there may be many more (or to some users, even better) results on lower positions.

- Every search engine faces a conflict of interest when it also acts as a content provider and shows results from its own offerings on its results pages (for example, Google showing results from its subsidiary YouTube). This problem gets exacerbated when one search engine has a large market share, as it is able to increase both its influence on its users as well as its suppression of its competitors' offerings.

- The more users rely on a single search engine, the higher the influence of search engine optimization (SEO) on the search results, and therefore, on what users get to see from the

It has been found that search results simply should not be considered "neutral."

Web. The aim of SEO is to optimize Web pages so they get ranked higher in search engines (that is, influencing a search engine's results). Taken together with the fact that SEO is now a multibillion-dollar industry,¹² we can see huge external influences on search engine results.

A Lack of Plurality

Considering these three problems, we can see that in the current market situation, we are far from plurality, not only in terms of the numbers of search engine providers but also in the number of search results. A Yahoo 2011 study showed that while we can regard a search engine as a possible window to all of the Web's content, more than 80% of all user clicks were found to go to only 10,000 different domains.⁴ We can assume these numbers are comparable for other search engines. Taken together, search engines have a huge influence on what we as users get to see on the results pages, and consequently, what we select from.

Why Are There No Alternatives?

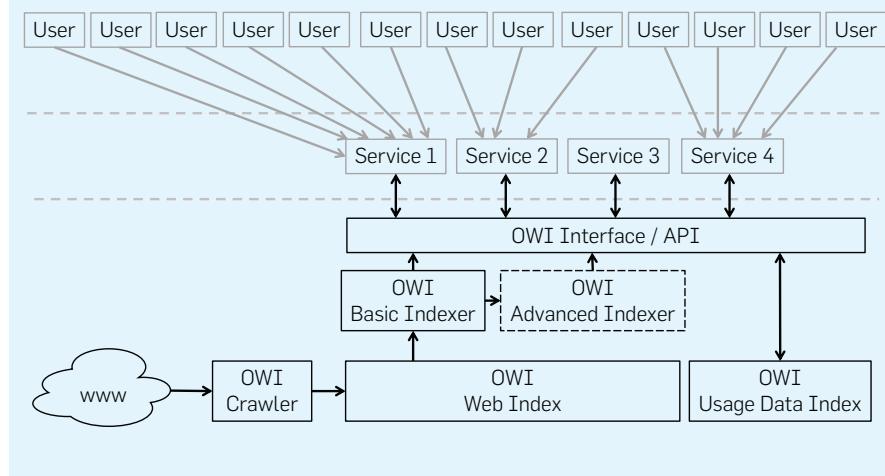
Why are there no real alternatives to the few popular search engine index providers? Firstly, index providers face huge technical difficulties due to the large numbers of documents resulting from the ever-changing nature of the Web. A second, significant, issue is the cost of hardware, infrastructure, maintenance, and staff. Thirdly, the Web is huge, and a search engine index needs to be tasked with covering as large a part of it as possible. While we know that no search engine can cover the Web in total, modern search engines know of trillions of existing pages.⁹ And indexing these pages is only the start. A search engine must keep its index current, meaning it needs to update at least a part of it every minute. This is an important requirement that is not being met by any of the current projects (such as Common Crawl) aiming at indexing snapshots of (parts of) the Web.

Separating Index and Services

I am proposing an idea for a missing part of the Web's infrastructure, namely a searchable index. The idea is to separate the infrastructure part of the search engine (the index) from the services part, thereby allowing for a multitude of services, whether existing as search engines or otherwise, to be run on a shared infrastructure.

The accompanying figure shows how the public infrastructure is responsible for crawling the Web, for indexing its content, and for providing

Separating services from infrastructure.



Distinguished Speakers Program

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

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

speakers.acm.org



Association for Computing Machinery

an interface/API to the services built upon the index. The indexing stage is divided between basic indexing and advanced indexing. Basic indexing provides the data in a form that services built on top of the index can easily and rapidly process that data. So, while services are allowed to do their further indexing to prepare documents, some advanced indexing is also provided by the open infrastructure. This provides additional information to the indexed documents (for example, semantic annotations). For this, an extensive infrastructure for data mining and processing is needed. Services should, however, be able to decide for themselves to what extent they want to rely on the preprocessing infrastructure provided by the Open Web Index. A design principle should be to allow services a maximum of flexibility.

As modern search engines rely heavily on usage data, this data (most prominently search queries routed to the index) is collected and made available for reuse. The OWI Usage Data Index allows for this data to be collected, stored, and queried. So, while each service can collect and query its own usage data, every service that wants to access usage data from the OWI Usage Data Index should be required to share anonymized usage data with the other services, so that every service profits from the amassed data. It is clear that existing search engines like Google and Bing have a huge lead compared to new providers, as they have a solid user base and already amassed large amounts of usage data. However, sharing usage data between the services could at least lessen the cold start problem.

Benefits

The main benefit of such an index would be for all interested parties to be able to develop their own applications without the problem of having to create their own index of the Web, which currently is an impossible endeavor not only, but especially, for small- and medium-size enterprises, as well as for non-commercial bodies.

Given a considerable uptake for such an index, it would foster plurality not only in the use of Web content

by developers but also in the variety of content that users get to see. We can rightly assume each search engine using the index would apply its own ranking function, and therefore, produce different results. Users would benefit in that they would not have to rely on only one or at best a few search engines but could choose from a variety of engines serving their different purposes. In that way, an Open Web Index would foster plurality and restrict the power of single companies dictating which content is shown to and consumed by users.

Another benefit would be that the index would be open to everyone, and therefore, would allow for investigating its transparency. However, search engines built on top of the index could still be "black boxes" in that they would not need to make their ranking functions open to anybody.

Possible Applications

While the Open Web Index would first and foremost make the development of new Web search engines feasible and financially attractive, it could also form the basis for a variety of other applications, being related to search or not.

In the field of search, the Open Web Index would also allow for vertical search engines (like image search, video search, or search in specific areas and on specific topics) to be built. In vertical search applications, OWI data could also be used to amend proprietary data. For instance, a provider of company information could amend its company profiles with Web data.

Apart from search, the OWI could also build the basis for data analysis

The index would be open to everyone, and therefore, would allow for investigating its transparency.

and topic detection and tracking. Examples of applications are opinion-mining tools and market research applications.

In the field of artificial intelligence, the Open Web Index could be used as a basis for large-scale machine learning. Likely applications in this area are machine translation, question-answering, and conversational applications.

Last but not least, an Open Web Index would provide a rich data source for researchers in many different fields, ranging from computer science and computational linguistics to computational social sciences and research evaluation.

It is clear this short list of ideas is far from being complete and only serves illustrative purposes. It shows, however, the huge potential of making Web data open to all parties interested.

Alternative Approaches

Some alternative solutions have been proposed for fostering plurality in the search engine market. The first and probably most obvious solution is to wait for commercial market players to develop alternatives. However, as we have seen in the last 15 years or so, Bing has been the only search engine capable of gaining considerable market share. Other search engines have failed, have been acquired by larger search companies, or have focused on niche markets. All new search engine providers face the problem of having to build their own index, which is, as has been described earlier, a very costly undertaking. Furthermore, what would be gained if we had one or two, even three more search engines on the market? From my point of view, the problem lies not in having a few more search engines, but in providing real search plurality.

The second line of argumentation says Google should be forced to provide fair and unbiased results. This is what the European Commission's competitive investigation against Google has been all about. However, as ranking results are always based on interpretations (and human assumptions inherent in the ranking algorithms), there is no such thing as an unbiased result set. Only a multitude of different algorithmic interpretations can help bring about search plurality.

Those that benefit from the index should have their say in building it.

The third line of argumentation calls for Google to open its index to third parties. Then, it would be possible to build (search) applications on top of Google's index. However, the control over the index—and over what third parties would be able to get from the index—would still lie in the hands of a private company, the index would still not be transparent, and there would still be no influence on how the index is composed.

The fourth, and already widely discussed solution, is building a publicly funded search engine as an alternative to the commercial enterprises. However, this again would only add one more search engine to the market, instead of fostering plurality.

Conclusion

The main idea I presented in this Viewpoint is to foster building search engines and other services needing Web data on top of a public infrastructure that is open to everyone. A multitude of such services would foster plurality not only on the search engine market (with the result of having more than a few search engines to choose from) but even more importantly, a plurality with regard to the results users get to see when using search engines.

Search results as a basis for knowledge acquisition in society seem too important to be left solely in the hands of a few commercial enterprises. The Open Web Index is comparable to other public services such as constructing roads and railroad tracks, supporting public broadcasting and, most notably, building a library system. An Open Web Index could be one of the main building blocks of the library of the 21st century.

An Open Web Index is a project that cannot and should not be un-

dertaken by a single company or institution. On the contrary, I envision building such an index as a task of society and for society, meaning we should build the index involving all actors and interest groups relevant to society at large. Those that benefit from the index should have their say in building it.

A question that remains is funding. As a considerable amount of money is needed, I argue for public funding not by a single state, but rather by a larger entity such as the European Union. This, however, does not mean a governmental body should also be the operator of the Open Web Index. Rather, it should be run by an organization that is relatively free from state intervention. One could think of a foundation running it or a model similar to public broadcasting. Whatever the mode of operation, as a project of and for society, funding should be applied for the greater good. □

References

1. Broder, A. A taxonomy of Web search. *ACM SIGIR Forum* 36, 2 (2002), 3–10.
2. Epstein, R. and Robertson, R.E. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. In *Proceedings of the National Academy of Sciences* 112, 33 (2015), E4512–E4521.
3. European Commission. Antitrust: Commission fines Google €2.42 billion for abusing dominance as search engine by giving illegal advantage to own comparison shopping service—Factsheet, 2017; <https://bit.ly/2tRknDj>.
4. Goel, S. et al. Anatomy of the long tail: Ordinary people with extraordinary tastes. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ACM (2010), 201–210.
5. Grimmemann, J. Some skepticism about search neutrality. *The Next Digital Decade: Essays on the Future of the Internet* 31, (2010), 435–460.
6. Lewandowski, D. Is Google responsible for providing fair and unbiased results? In M. Taddeo and L. Floridi, Hrsg., *The Responsibilities of Online Service Providers*. Springer, Berlin Heidelberg, 2017, 61–77.
7. Noble, S.U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, NY, USA, 2018.
8. Purcell, K., Brenner, J., and Raine, L. *Search Engine Use 2012*. Washington, D.C., USA, 2012.
9. Schwartz, B. Google's search knows about over 130 trillion pages. *Search Engine Land*, 2016; <https://selnd.com/2g7MnA7>.
10. Sterling, G. Data: Google monthly search volume dwarfs rivals because of mobile advantage. *Search Engine Land*, 2017.
11. Sullivan, D. Google now handles at least 2 trillion searches per year. *Search Engine Land*, 2016; <https://selnd.com/2GsdYYq>
12. Sullivan, L. Report: Companies will spend \$65 billion on SEO in 2016. *Media Post*, 2016; <https://bit.ly/2BqNrqX>
13. White, R.W. and Horvitz, E. Cyberchondria. *ACM Transactions on Information Systems* 27, 4 (2009), Article No. 23.

Dirk Lewandowski (dirk.lewandowski@haw-hamburg.de) is Professor for Information Research and Information Retrieval at the Hamburg University of Applied Sciences in Hamburg, Germany.

Copyright held by author.

Europe Region Special Section

THE COMPUTING COMMUNITY throughout the European Region is championing many enterprising industry, academic, and government initiatives to further develop the field and ensure a workforce prepared to take it on.

The articles in this special section, written by some of the leading voices in the region, tell stories of informatics and ICT innovations, Web science in Europe, the EuroHPC, future research directions planned for this vibrant part of the world, and much more.



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



Welcome to the Europe Region Special Section

WITH ITS POPULATION of over 740 million people and 24 official languages, Europe provides a unique environment for the development of a distinctive computing landscape. We believe you will see this reflected in this first Europe Region Special Section.

We invited a mix of practitioners and academics from across the Europe region, not only European Union members but also Switzerland and Israel, to suggest topics for inclusion in this Europe Region Special Section. We brainstormed article ideas with nearly two-dozen colleagues from across the region at a workshop in Paris in July 2018. The article pitches generated at that workshop went through two further rounds of review and refinement. As you will see, the resulting collection of articles offers an excellent view of some of the most exciting activities in the region.

We are proud to present this section which, like the China Region Special Section published last November, consists of two kinds of articles: shorter “Hot Topics” articles that set the context for European developments in computing, and longer “Big Trends” articles that describe exciting developments in areas such as high-performance computing, embedded systems, and computing education. Half of the Hot Topics articles describe the European consumer computing market, the innovation ecosystem, demographics, and the research agenda; the remaining articles briefly describe specific initiatives and recent developments.

Europe has played an important role in the development of the computing discipline, from the pioneering work of Alan Turing and Konrad Zuse through the modern day. The following pages show that it continues to be a vibrant, distinctive part of the global computing community. We are sure the reader will find much of interest in the following pages.

— **Panagiota Fatourou and Chris Hankin**
Europe Region Special Section Co-Organizers

Panagiota Fatourou is an associate professor in the department of computer science of the University of Crete, Greece, as well as a collaborating faculty member at the Institute of Computer Science of the Foundation for Research and Technology, Hellas (FORTH).

Chris Hankin is co-director of the Institute for Security Science and Technology, and a professor of computing science, at Imperial College London in the U.K.

Copyright held by owners/authors.



Back row from left: Yota Papageorgiou, Julie McCann, Steven Newhouse, Joaquim Jorge, Koen De Bosschere, Guillaume Toublanc, Michael Caspersen, Chris Hankin. **Front row from left:** Fabrizio Gagliardi, Paola Inverardi, Andrew Chien, Panagiota Fatourou, Vassiliki Petousi, Lihan Chen.

EDITORIAL BOARD

EDITOR-IN-CHIEF

Andrew A. Chien
eic@cacm.acm.org

DEPUTY TO THE EDITOR-IN-CHIEF

Lihan Chen
cacm.deputy.to.eic@gmail.com

CO-CHAIRS, REGIONAL SPECIAL SECTIONS

Sriram Rajamani
Jakob Rehof

EUROPE REGION SPECIAL SECTION CO-ORGANIZERS

Panagiota Fatourou
University of Crete
Chris Hankin
Imperial College London



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

Hot Topics



Big Trends



- 32 A Demographic Snapshot of the IT Workforce in Europe**
By Lisa Korrigane
-
- 34 Enterprises Lead ICT Innovation in Europe**
By David Pringle
-
- 35 Europe's Ambitious ICT Agenda**
By David Pringle
-
- 36 Europe's Well-Connected Consumers**
By David Pringle
-
- 38 New European Data Privacy and Cyber Security Laws—One Year Later**
By Laurence Kalman
-
- 40 Incorporating Europe's Values in Future Research**
By Jan Gulliksen
-
- 42 HiPEAC: A European Network Built to Last**
By Koen De Bosschere, Marc Duranton, and Madeleine Gray
-
- 44 ACM Europe Council's Best Paper Awards**
By Joaquim Jorge, Mashhuda Glencross, and Aaron Quigley

- 46 Connected Things Connecting Europe**
By Julie A. McCann, Gian Pietro Picco, Alex Gluhak, Karl Henrik Johansson, Martin Törngren, and Laila Gide
-
- 52 Women Are Needed in STEM: European Policies and Incentives**
By Panagiota Fatourou, Yota Papageorgiou, and Vasiliki Petousi
-
- 58 Informatics as a Fundamental Discipline for the 21st Century**
By Michael E. Caspersen, Judith Gal-Ezer, Andrew McGetrick, and Enrico Nardelli

- 64 The European Perspective on Responsible Computing**
By Paola Inverardi
-
- 70 Toward a European Exascale Ecosystem: The EuroHPC Joint Undertaking**
By Thomas Skordas
-
- 74 Web Science in Europe: Beyond Boundaries**
By Steffen Staab, Susan Halford, and Wendy Hall

Workforce | DOI:10.1145/3309915

A Demographic Snapshot of the IT Workforce in Europe

BY LISA KORRIGANE

EUROPE IS NOT a static entity but here is what it looks like in 2019:

The European Union is made up of 28 countries. The capital is in Brussels, Belgium, and the presidency is shared among EU members each semester. In 2019, the first semester sees Romania holding presidency until June, then Finland until the end of the year. An estimated 551.8 million people live in the EU, speaking 24 official languages. Approximately

72% of the population is employed,^a which is greater than the pre-economic-crisis peak of 2008. Men are more employed than women by an average of 11%.^b

The ICT workforce in the EU counts some 8.4 million people. The U.K. alone provides almost 20% of this workforce, although it only accounts for 12.8% of the EU population. The second and third providers are Germany and France, but the propor-

a <http://bit.ly/2CGkreW>
b <https://bit.ly/2RYTmNI>

tions are more coherent with their population ratios. The number of ICT specialists has grown over 36% in the last 10 years,^c a significant jump from a mere 3.2% a decade ago. This marked increase helps explain why ICT employment has resisted the effects of the region's overall financial crisis.

Obviously, the raw numbers favor the most populated countries, but the proportion of ICT specialists in total employment country by country tells a different story, as Nordic countries lead the way. Finland comes first with 6.8% of its workforce dedicated to ICT.^d In comparison, Germany is much further behind with 3.8%, and France tailing Germany with 3.7%. The other two leaders are Sweden and Estonia with 6.6% and 5.6% of their respective workforces dedicated to ICT. Both Finland and Sweden

are home to very efficient teaching methods as they integrated computer studies into their school curricula since the Nokia and Ericsson years. The two mobile phone companies actually spirited a whole generation to place their trust in ICT. As for Estonia, political decisions taken 20 years ago turned the country into "one of the most advanced digital societies in the world."^e Despite its small size, Denmark is also very active in ICT and has attracted big corporate names, such as Microsoft, Uber, and IBM.

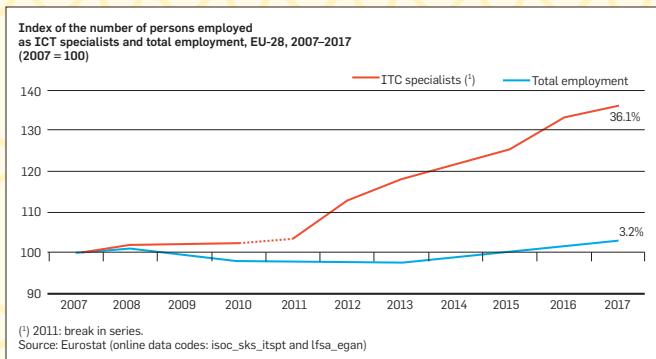
The countries that lead the ICT sector in Europe today are the ones that invested a great deal of time and resources 20 years ago, especially in education.

The Typical ICT Worker and the Gender Gap

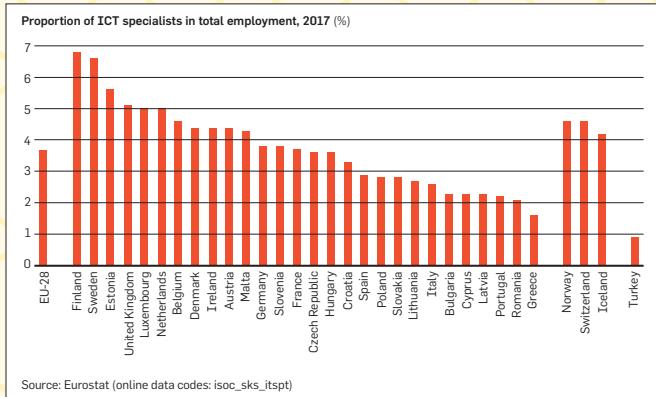
The overall picture of the European ICT worker can be summed up as follows: A male over 35 with tertiary education diploma. Almost two-thirds of all people employed as ICT specialists in the EU are over 35 years of

c <http://bit.ly/2Dr16zI>
d <https://bit.ly/2S3Lfzo>

e <https://e-estonia.com>



Index of the number of persons employed as ICT specialists.



Proportion of ICT specialists in total employment.

age.^f Europe does have quite a significant unemployment rate of 65.3% for those under 24 years old. Postsecondary education also plays a role in the ICT population as over 62% of all ICT specialists in the EU have completed tertiary level education,^g with the highest shares of this attainment found in Lithuania, Ireland, Cyprus, and Spain.

The vast majority of people employed as ICT specialists are men. They account for 82.8% of the total ICT workforce.^h The number of women in ICT has actually decreased by 5% during the last 10 years, with only slight increases noted in France, Belgium, and the Netherlands. However, Bulgaria has the highest proportion of women in ICT as they account for 26.5% of the workforce. The total

proportion of employed women in the EU (across all sectors) is 66.4%, but rises well above those figures in countries like Germany (75.2%), Estonia (75.1%), and Finland (72.4%).

The dear need of ICT specialists. The demand for ICT specialists in the EU is very high. One in five businesses need ICT workers across all sectors of the economy.ⁱ Larger companies have also reported challenges recruiting skilled ICT workers. In 2017, 48% of nine companies that recruited or tried to recruit reported difficulties in filling vacancies. This applies to all countries in the EU. Depending on projections,^j the expected shortage of 10 skilled ICT workers in the EU starts at 526,000 individuals in 2020. In the case of a high-growth estimate, this shortage could

reach over 900,000 workers. Because the shortage is global, salaries have increased to attract skilled individuals. Salaries in the U.S., however, are much higher.^k For example, the average salary for a software developer in the U.S. is \$92,240 and \$43,749 in France. This induces a brain drain of talent, which brings about two pernicious side effects in Europe: lack of senior specialists and lack of qualified trainers for upcoming workers. However, a mere comparison of salaries would be inaccurate. Workers in France pay very little for healthcare and have almost no education expenses (both financed with taxes). Most young workers start their professional life with no student loans to pay back.

The sectors where demands are the highest vary according to studies, but all agree that big data analytics and business analytics^l are the most sought-after.

Norway, Switzerland, Iceland, and Turkey are not members of the EU, but are located inside or close to the European continent. Is their position in the ICT ecosystem any different from EU members? Norway, Switzerland, and Iceland relate to the EU's average statistics. Turkey, on the other hand, sets itself apart with just 0.9% of its workforce in ICT and only 10% of those are women.^m Moreover, Turkey has almost two-thirds of its ICT workforce under the age of 34, which is the exact opposite of the EU numbers.

Europe's Efforts to Meet Demand for ICT Specialists

Despite a variety of ICT

opportunities available in Europe, it lags behind the U.S. and China. The brain drain mentioned earlier and the difficulty to standardize actions throughout the vast continent hinders efforts to catch up.

Following the demands of top IT scientists in 2018, the European Laboratory for Learning and Intelligent Systems (ELLIS) was founded last December.ⁿ Focusing mainly on artificial intelligence and, more broadly, on machine learning, they aim to create a network to advance breakthroughs across the continent and educate the next generation of AI researchers.

The U.K. invested £1 billion in artificial intelligence^o in 2018 and created the Centre for Data Ethics and Innovation to monitor the AI research.

The countries that lead the ICT sector in Europe today are the ones that invested a great deal of time and resources 20 years ago, especially in education. In order to catch up, much effort must be focused on developing computer science in school.

It appears that a huge potential ICT workforce resides with women. They must be encouraged, very early on in school, to embrace ICT careers. This would lead to a more balanced sector, improve women's employment rates, and help reduce the shortage in ICT specialists.^p □

^f <http://bit.ly/2S3dJjt>

^g <http://bit.ly/2CAWY8f>

^h <http://bit.ly/2MpOSuj>

ⁱ <http://bit.ly/2S8aReB>

^j <http://bit.ly/2FSIaNi>

^k <http://bit.ly/2DqNRPI>

^l <http://bit.ly/2FIkhYj>

^m <http://bit.ly/2MpOSuj>

ⁿ <http://bit.ly/2FJZ9Re>

^o <http://bit.ly/2FIY2kN>

^p <http://bit.ly/2FSIzIM>

Lisa Korrigane is a freelance writer, reporting mostly on the technology market, startups, and digital policies in France. She is currently studying to become a Web developer.

Enterprises Lead ICT Innovation in Europe

BY DAVID PRINGLE

IN GLOBAL TERMS, Europe's information and communications technology (ICT) industry is small, overshadowed by the massive U.S. software industry and the extensive electronics industry in East Asia. But it does have some world-class companies, such as telecom equipment providers Nokia and Ericsson, online music platform Spotify, e-commerce company Zalando, enterprise software provider SAP, games developer Supercell (now owned by Tencent), embedded processor designer ARM (now owned by Softbank), and Skype (now owned by Microsoft). Moreover, some of the region's telecom groups, Deutsche Telekom, Vodafone, Orange, and Telefónica, are major multinationals with operations spanning several continents.

Although it has only one of the top 10 artificial intelligence (AI) research institu-



Laura Citron, chief executive of London and Partners, welcomes technology talent from around the world to the 2018 Deep Tech Summit.

tions^a (CNRS in France) and no major cloud service providers or Internet platforms on the scale of Amazon, Microsoft, or Google, Europe does innovate in ICT. Non-tech companies, particularly automakers, banks, and pharmaceutical firms, such as BMW, Deutsche Bank, BNP, and Bayer, drive much of this innovation. Whereas

^a Ranked by most cited AI-related research papers. The Nikkei & Elsevier, Atomico State of European Tech 2017; <https://2017.stateofeuropeantech.com/>

Europe's tech industry is cash-strapped, non-tech companies in Europe have more cash holdings than their counterparts in the U.S. or China.^b

As one of the world's leading financial centers, London hosts a thriving fintech industry, which is harnessing big data analytics to personalize financial services. Germany and France are home to many of the world's leading players in transportation, which are adopting ICT rapidly, as they move to semi-autonomous vehicles and ultimately self-driving cars. BMW, Daimler, Siemens, Bosch, and Airbus are among the major European companies embracing the Internet of Things to improve trucks, trains, planes, and cars. For example, the region is leading trials of platooning—the use of wireless technologies to enable semi-autonomous trucks to drive

^b The S&P CapitalIQ Platform. Atomico State of European Tech 2017; <https://2017.stateofeuropeantech.com/>

"Europe is made of tens of different cultures, it's our biggest advantage. The more diversity you can find among entrepreneurs, the more you will find innovative businesses, creativity, and value created."

in convoys along motorways, while their drivers take a break. Platooning could cut fuel costs, reduce congestion, and increase efficiency.

Rising Investment in Deep Tech Research

At the same time, the region's tech ecosystem is renewing itself: In Europe, \$3.5 billion was invested in so-called deep tech companies in 2017 across more than 600 deals, up from \$2.5 billion in 2016.^c Deep tech refers to software, semiconductors, and other digital hardware. Many of these investments take advantage of Europe's renown computer science institutions—the continent is home to half of the top 10 computer science institutions in the world.^d Moreover, Europe is producing more than twice the STEM Ph.D.'s graduating in the U.S.^e

AI is the hottest area for investment. AI companies in Europe have raised more than \$4.6 billion since 2012 across over 1,000 deals. Europe has also spawned several hundred blockchain companies and augmented reality or virtual reality startups since 2012.^f Helsinki has the second highest concentration of app developers in the world behind the San Francisco Bay area.^g Minsk in Belarus ranks sixth, Stock-

^c Dealroom.co. Atomico State of European Tech 2017; <https://2017.stateofeuropeantech.com/>

^d Times Higher Education World University Rankings 2018. Atomico State of European Tech; <https://2018.stateofeuropeantech.com/>

^e The OECD. Atomico State of European Tech 2017; <https://2017.stateofeuropeantech.com/>

^f Tracxn. Atomico State of European Tech 2017; <https://2017.stateofeuropeantech.com/>

^g A study by Caribou Digital; <http://bit.ly/2REupYi>

holm eighth, and London ninth on this measure.

"The level of capital available has skyrocketed in the past years," noted Xavier Niel of Station F, in an interview with Atomico. "Europe is made of tens of different cultures, it's our biggest advantage. The more diversity you can find among entrepreneurs, the more you will find innovative businesses, creativity, and value created."

In 2015, the value added by the EU ICT sector amounted to \$654 billion, 5.2% more than the previous

year.^h The sector employed 6.4 million people and spent \$34 billion on business R&D expenditure.

Major regional disparities in pay. Yet, Europe still sees many of its entrepreneurs and ICT specialists emigrate to North America, either to obtain funding, expertise, or higher pay. On average, software developers in Switzerland earn \$85,709—more than anywhere else in Europe, but less than their peers in the U.S. (\$92,240). Norway

^h The European Commission, <https://bit.ly/2sGADYP>

is next with \$70,776, followed by Denmark (\$70,082), the U.K. (\$59,268) and Germany (\$57,345).ⁱ In fact, there are major regional disparities. Whereas the median salary for a software engineer in Berlin, London, and Paris is over \$50,000, the equivalent figure in Warsaw is less than half that and just \$15,000 in Bucharest. By way of comparison, a software engineer in Tokyo, Japan, earns almost \$54,000 on average.^j

However, by some mea-

ⁱ Daxx.com; <https://bit.ly/2FKh7Dl>
^j Glassdoor; <https://bit.ly/2RKEKBZ>

sures, Europe is better than other regions at harnessing its talent. Indeed, some of Europe's education systems are the envy of the world: Estonia and Finland are in the top five globally, while 13 European countries rank above the U.S. in terms of education outcomes, according to the PISA 2015 study of 15-year-olds' performance in math, science, and reading. □

David Pringle is a London-based writer for Science|Business Publishing Ltd., covering the telecom, media, and technology sectors.

© 2019 ACM 0001-0782/19/4 \$15.00

ICT Plans | DOI:10.1145/3309919

Europe's Ambitious ICT Agenda

BY DAVID PRINGLE

FOR EUROPE, INVESTMENT in advanced ICT is a must. With an aging population and a shrinking workforce, Europe needs to tap artificial intelligence (AI), 5G wireless connectivity, quantum computing, and other ICT technologies that could drive the next step change in productivity.

To that end, the region can build on a long-standing scientific tradition. Thanks in part to sustained public sector support, Europe is a leading producer of high-quality scientific research. Its scientists excel in aeronautics, transport technologies, and energy and construction, based on the number of widely cited publications.^a



Leading European AI researchers assembled in Montreal last December to announce the establishment of a society to found a cross-national European Laboratory for Learning and Intelligent Systems (ELLIS).

As a densely populated continent, Europe is at the forefront of urban planning and the development of smart transport systems. London's \$20 billion Cross-rail project, which plans to introduce new intelligent on-train management systems to reduce energy costs by up to 20%, is one of many examples of Europe's

commitment to public transportation.

And Europe remains a leader in the automotive sector,^b where major

^a The European Commission, Science, Research and Innovation Performance of the EU (SRIP) report: <http://bit.ly/2DsKasE>

carmakers and well-funded start-ups are driving a shift toward electric propulsion and self-driving systems. In Sweden, for example, Northvolt has broken ground on a \$4.2 billion factory to produce lithium ion batteries for electric cars.

Indeed, European policymakers reckon batteries will be as essential to the automotive industry of the 21st century as the combustion engine was in the 20th century. The European Commission (EC), the European Investment Bank, and over 260 industrial and innovation stakeholders have joined the European Battery Alliance (EBA), which is now building its first pilot production facilities. The EC estimates Europe will need at least 20 'gigafactories' (large-scale battery cell production facilities) to meet local demand.

Building a CERN for AI

But maintaining its industrial and manufacturing base may require Europe to raise its game in AI and robotics, which promise to drive another industrial revolution. To that end, leading European scientists are trying to establish the European Lab

for Learning and Intelligent Systems (ELLIS), a multinational institute that would be devoted to AI research. The concept is modeled on CERN, the particle physics lab created after the World War II to stem the flow of physicists across the Atlantic. Although it is not clear whether ELLIS will get off the drawing board, the EC has promised to spend an additional \$1.7 billion on AI research between 2018 and 2020, which it hopes will stimulate a further \$2.8 billion investment by public-private partnerships.

That comes on top of a pledge by the EC and EU members states to spend \$1.1 billion building world-class supercomputers, after recognizing that Europe is also falling behind in this area. "We do not have any supercomputers in the world's top 10," Andrus Ansip, EC Vice-President for the Digital Single Market, said in January 2017. "We want to give European researchers and companies world-leading supercomputer capacity by 2020—to develop technologies such as artificial intelligence and build the future's everyday applications in areas like health, security or engineering." However, both China and the U.S. are also investing heavily in AI research and supercomputing capacity.

Commercializing Quantum Computing and 5G

The EC is also anxious for Europe to commercialize quantum computing. Blogging about its new \$1.1 billion Quantum Flagship initiative, Ansip wrote: "While Europe has many world-class scientists in this field, there is so far little industrial take-up or commercial exploitation here." After

the Graphene Flagship and the Human Brain Project, the Quantum Flagship is the third large-scale research and innovation initiative of this kind funded by the EC.

At the same time, Europe wants to lead the development and deployment of 5G wireless technologies. In 2018, the non-profit European Investment Bank has lent \$580 million to Nokia and \$300 million to Ericsson to further R&D related to 5G. To help build a global consensus on future 5G standards and spectrum requirements, the European Commission has established Joint Declarations on 5G with Brazil, China, Japan, and South Korea. Cooperation agreements are also being discussed with India and the U.S.

Trailblazing on Global Collaboration and Regulation

With 50 countries packed into one continent, Europeans are well accustomed to international collaboration, as evident in its major strategic alliances, such as Airbus, CERN, the European Molecular Biology Lab, and the European Space Agency, which are all backed by multiple countries within Europe. European businesses also tend to be supportive of international standards: Europe was the birthplace of GSM, the technology that brought mobile communications to the world.

Furthermore, Europe's regulators are highly influential. EU directives

and regulations often form the basis for government interventions in other markets. The General Data Protection Regulation (see Laurence Kalman's article on p. 38) and the second Payment Services Directive, which both came into force in 2018, provide consumers with sweeping new rights to protect and extract their personal data. This radical legislation, together with the multibillion-dollar fines levied on Google, underlines the EU's readiness to try and exert more control over disruptive digital players from outside its borders. □

David Pringle is a London-based writer for Science|Business Publishing Ltd., covering the telecom, media, and technology sectors.

© 2019 ACM 0001-0782/19/4 \$15.00

Consumers | DOI:10.1145/3309921

Europe's Well-Connected Consumers

BY DAVID PRINGLE

HOME TO APPROXIMATELY 740 million people, many of them affluent, Europe spends a lot of money on information and communications technology (ICT). The European ICT market was worth \$769 billion in 2017 (up 1.8% from 2016).^a

Yet, despite the best efforts of the European Union (EU), Europe is not one market. There are major cultural differences and economic disparities between northwest Europe and southeast Europe. Whereas Germany, the U.K., the Nordics, and

the Netherlands tend to attract migrants from all over the world, many countries on the eastern and southern rims of Europe are seeing an exodus of young people and low birth rates. Indeed, the continent as a whole is aging: One fifth of the people in the 28 members of the EU (the EU28) are now 65 or over, compared with 17% in 2007.^b In the U.S., the equivalent figure is 15% and in China 11%.^c

The vast majority of Europeans are online. It is relatively cost-effective

for the region's telecoms companies to provide connectivity: Europe is densely populated and heavily urbanized—three quarters of the EU population lives in cities, towns, or suburbs. Across the EU28, more than 87% of households had Internet access and 85% broadband Internet access at the end of 2017.^d Moreover, the broadband is relatively quick: Of the top 50 countries ranked by broadband speeds worldwide, 36 of them are in Europe.^e Sweden has the fastest broadband in Europe, offering an average

^a Eurostat; <http://bit.ly/2TaeXQs>

^b The World Bank: <https://data.worldbank.org/indicator/SP.POP.65UP.TO.ZS>

^c Eurostat; <http://bit.ly/2sHEpB8>

^d Comparison website cable.co.uk; <http://bit.ly/2DsPBYH>

speed of 46Mbps. However, interference between Wi-Fi networks is common in the many districts where people live in apartment buildings, while cellular networks can also be heavily congested in city centers.

Most Europeans now have smartphones. Approximately two-thirds of people in the EU28 between ages 16 and 74 had mobile Internet access at the end of 2017, up from 36% in 2012. But there are wide regional variations, with that figure reaching 87% in the Netherlands and Sweden, compared with just 32% in Italy and 40% in Poland. In 2017, close to three quarters (72%) of individuals in the EU28 accessed the Internet on a daily basis, with a further 8% using it at least once a week.

A Battleground for North American and East Asian Technologies

Lacking a major computing industry of its own, Europe is a relatively neutral market for hardware and software made in both North U.S. and East Asia. Major American and Asian brands go head-to-head in the smartphone, tablet, and computing markets, but their operating systems all hail from the U.S. Android dominates the European smartphone market. Some 70% of smart-

phones in use in Europe run Android, while 28% run Apple's iOS, while less than 1% run Windows.^f However, in the tablet market, iOS has a market share of 66%, while Android has 34%. But there is one key smartphone component that hails from Europe—U.K.-based ARM Holdings' microprocessor architecture is used in more than 90% of the world's handsets. This low-power architecture has proven pivotal in the development of advanced handsets with long battery lives.

Social networking in many European countries is not as prevalent or as popular as in North America, the birthplace of Facebook and other leading social networks. Just over half (54%) of Europeans age 16 to 74 use the Internet for social networking, while in France and Italy that proportion is as low as 43%,^g potentially reflecting a preference for face-to-face interactions. In the same demographic, 57% of Europeans shop online, while 18% are using accommodation-sharing services, such as Airbnb, and 8% use

^f Statcounter; <http://gs.statcounter.com/os-market-share/mobile/europe>

^g Eurostat (figures for end of 2017); <http://bit.ly/2sHEpB8>



IMAGE BY RADIOKAFKA/SHUTTERSTOCK.COM

Europeans tend to be greener than North Americans. More than nine in 10 respondents (94%) say that protecting the environment is important to them personally.

ride-sharing services, such as Uber.

In Europe's three largest economies (France, Germany, and the U.K.) YouTube and Netflix are the top video streaming apps, ahead of local media players, while WhatsApp Messenger, Facebook, and Facebook Messenger are the top three social apps in these markets.^h Although major U.S. Internet services are widely used across Europe, some smaller players also have significant traction. For example, London-based music recognition service Shazam, which was recently acquired by Apple, is ranked sixth in Italy in terms of monthly active users, and seventh in France, and ninth in Spain. In Russia, a market apart, cultural and regulatory factors have helped several local players, including Yandex, Mail.Ru, and Sberbank, compete very effectively with the global players. All of these have apps in Russia's top 10, as ranked by monthly active users.

Europeans Care about Privacy and the Environment

Privacy is a big deal for Europeans, particularly in Germany, which is very wary of state surveillance after the country's experience of authoritarianism in the

^h App Annie; <https://www.appannie.com>

first half of the 20th century: Some 45% of Europeans who use the Internet have installed or changed their antivirus software in the past three years due to privacy and security issues, while 39% say they are now less likely to share personal information on websites.ⁱ More than six in 10 respondents (61%) say the security and privacy features of an IT product play some role in their choice, while 27% are ready to pay more for better security and privacy features.

Europeans also tend to be greener than North Americans. More than nine in 10 respondents (94%) say that protecting the environment is important to them personally, and among these more than half (56%) say it is very important.^j These findings have remained broadly consistent over the past decade.

Europeans have mixed feelings about the direction in which ICT is headed. Although more than six in 10 respondents have a positive view of robots and artificial intelligence, an even higher proportion (72%) agree robots and AI steal jobs. □

ⁱ Eurobarometer; <http://bit.ly/2FPIlb0>

^j Eurobarometer; <http://bit.ly/2RIC47V>

David Pringle is a London-based writer for ScienceBusiness Publishing Ltd., covering the telecom, media, and technology sectors.

© 2019 ACM 0001-0782/19/4 \$15.00

New European Data Privacy and Cyber Security Laws—One Year Later

BY LAURENCE KALMAN

THIS HAS BEEN a momentous year for data protection and information security regulation in Europe, with two landmark pieces of legislation taking effect. Together they represent a major shift in the European industry's approach to privacy and security compliance.

The long-awaited General Data Protection Regulation (GDPR) came into force in the European Union (EU) on May 25, 2018, attracting a huge amount of attention and prompting a flurry of email messages to customers on historic marketing lists.

Now organizations that process personal data are regulated not only if they are established in the EU, but if they target goods or services at, or monitor the behavior of, individuals in the EU—regardless of where they are located. Service providers that process personal data for others become directly



regulated, while individuals' rights to manage their data have been enhanced. And the new sanctions regime has given regulators real teeth, with the ability to levy fines up to the greater of €20 million or 4% of total worldwide annual turnover.

Shortly before the GDPR took effect, the deadline for EU member

states to implement the Network and Information Security Directive (NISD) passed much more quietly. Often viewed as the GDPR's 'younger sibling,' the NISD has proven a less eye-catching piece of legislation although it too threatens hefty penalties for breach.

Whereas the GDPR focuses on protecting individuals' rights to privacy, the NISD originates in national security concerns. It aims to raise levels of cyber security in specific sectors that represent 'critical national infrastructure,' such as energy, transport, health and water, as well as among suppliers of essential digital services.

A New Culture of Compliance?

The GDPR has pushed data privacy compliance up the corporate agenda for the long term. Organizations must understand and document the personal data they use in far greater detail than before. Shortly before the compliance deadline, the International Association of Privacy Professionals and Ernst and Young estimated that large British firms had spent \$1.1 billion on GDPR preparations, while U.S.-based companies had invested \$7.8 billion.

According to research into GDPR readiness costs among FTSE100 companies carried out by management

Many view Europe's approach to data privacy and cyber security as setting a global gold standard.

consultants Sia Partners, banks were the biggest spenders at over £60 million on average. Next came the energy, commodities and utilities, retail goods, and technology and telecommunications sectors, with an average implementation expenditure of approximately £15–£19 million per company. And ongoing obligations under the GDPR will create a lasting increase in compliance costs across sectors.

Compliance also plays a major role under the NISD. The regulatory burden represents a greater shock to the system for industries that have not previously been required to prioritize cyber security or incident reporting. Organizations are also spreading the NISD compliance burden along their supply chains into sectors that are not directly regulated. One beneficiary of the increase in compliance risk is the insurance sector, and both the NISD and GDPR will continue stimulating the cyber insurance industry.

Growing Consumer Awareness

The GDPR in particular has had a noticeable effect on improving individuals' awareness of, and assertiveness in exercising, their data privacy rights. Organizations that hold large volumes of customer data have received rising numbers of data subject access requests, which can be costly to comply with.

Greater consumer awareness is also evident in increasing levels of interaction with regulators. In the first few months after the GDPR came into effect, the French regulator reported a 64% increase in complaints from individuals, which

in its view showed that EU citizens had warmly embraced the regulation.

Alongside the introduction of the GDPR and NISD, the European Commission has emphasized that building a European data economy is a key part of its 'digital single market' strategy.

But there is a natural tension between the desire to protect data privacy, boost cyber security, and promote a burgeoning European economy based on free-flowing data. It remains to be seen whether a data-driven economy can continue to flourish once the new regulations really start to bite.

In this more hostile environment, businesses and regulators will need to work hard to avoid a situation where Europe becomes a less attractive region to test and roll out new products. Balancing the free flow of data with respect for privacy and security concerns will be essential to the success of a dynamic, connected European economy.

On the other hand, many view Europe's approach to data privacy and cyber security as setting a global gold standard. Numerous countries still have no coherent data protection laws in place at all. Only a select few—Andorra, Argentina, Canada, the Faroe Islands, Guernsey, Israel, the Isle of Man, Jersey, New Zealand, Switzerland, Uruguay, and the U.S. (only under the Privacy Shield framework)—have data protection laws that reach the required threshold to be considered adequate by the EU. China is moving toward stringent data protection standards but still has a patchwork of regulation in place.

There are also signs

There is a natural tension between the desire to protect data privacy, boost cyber security, and promote a burgeoning European economy.

that U.S. consumers look longingly at the protections available in the EU. According to a survey conducted in April 2018 by Janrain, the customer profile and identity management software provider, 68% of respondents wanted a GDPR-like law in the U.S. Some 38% identified their top priority as the ability to control how their data is used, while 39% focused on the right to require organizations to delete their data.

When Will the Other Shoe Drop?

The U.K. Information Commissioner's Office (ICO) issued its first formal GDPR notice in July 2018. This required data analytics firm AggregateIQ to stop processing data relating to U.K. individuals that it held through its work for the 'Leave' campaign in the EU membership referendum, and that it continued to process in breach of the GDPR. Following an appeal, the ICO narrowed its enforcement notice and no fine has been issued.

One of the first GDPR fines was issued by the Portuguese regulator against the Centro Hospitalar Barreiro Montijo in July 2018. A fine of €400,000 was levied against the hospital for two GDPR breaches relating to unauthorized access to patients' data and inadequate data security—

still a relatively modest amount in contrast to the maximum available. In January 2019, the French regulator raised the stakes by issuing a record €50m fine against Google, due to insufficient transparency, inadequate information, and a lack of valid consent in relation to personalized advertisements.

The full impact of the GDPR and NISD will therefore become clearer as regulators flex their muscles and issue more large-scale fines. Although we have not yet witnessed the predicted rush of group litigation, an uptick in data protection-related class actions is also likely.

The GDPR and the NISD are still in their relative infancy, but they will be with us for a long time to come. Generating trust will be the key to success in this increasingly connected world. Organizations must show they take cyber security and data privacy concerns sufficiently seriously to win consumers' confidence. Doing this while also providing market-leading services will enable Europe's data-driven economy to succeed in the years to come. □

Laurence Kalman is a partner at the London offices of CMS, an international law firm, where he works in the commercial/technology sector.

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

Incorporating Europe's Values in Future Research

BY JAN GULLIKSEN

OUR SOCIETY is currently undergoing several big changes that pose challenges and opportunities for the future. The increasing digitalization and automation, the growing globalization, and improved financial durability offer many excellent opportunities for development. There is more research funding in the system than ever before. On the other hand, our society is vulnerable; we have challenges in relation to inequality, an environment put under severe stress, and more hostile tendencies than we have seen in a long time, despite the good times and economic growth. In our work at the European Commission Independent High Level Group on Maximizing the Impact of EU Research & Innovation Programmes,⁴ we try to understand and

elaborate on these challenges to be able to propose a suitable strategy for future research funding from the European Commission.

With an area of 10 million square kilometers and a population of 740 million, Europe is a substantial region of prosperity and development. Many EU countries lead in rankings of prosperity, education, digitalization, equality, and low corruption.

Investment in research and innovation has been substantial and recognized as important for the development of the society, for ensuring a high level of skills, and for contributing to the creation of jobs and growth, albeit not to the extent demonstrated by North America or South East Asia. With just 7% of the world's population and 24% of global GDP, the EU produces approximately 30% of the world's scientific publications.⁴ Several European



countries are leading the research investment competition as a share of the country's GDP—after the three top countries Israel, Korea, and Japan—with Europe at 2.03% and Sweden leading in Europe with investments of 3.26% of the GDP.^a In the Lamy report,⁴ we argued for a European target of 3%.

However, the business economy in high-tech sectors and PCT patent applications per-million population both show a lower growth rate in Europe than in U.S. Hence, Europe must focus on innovation and investigate the mechanisms within its society that prevent development on a scale the same as the U.S.

The EU funding program

Horizon 2020 that focuses on scientific impact was prioritized particularly through the ERC program for funding excellent basic research. This turned Europe into an attractive arena to develop research careers and thus strengthening European research quality and performance.

Trends and Needs for Future

European Research

Some of the most prevalent trends and needs that may influence future research and innovation throughout Europe include the following:

- *Societal challenges of importance and acuteness.* The EC has conducted activities to foster mission-driven science and innovation based on the activities by Maz-

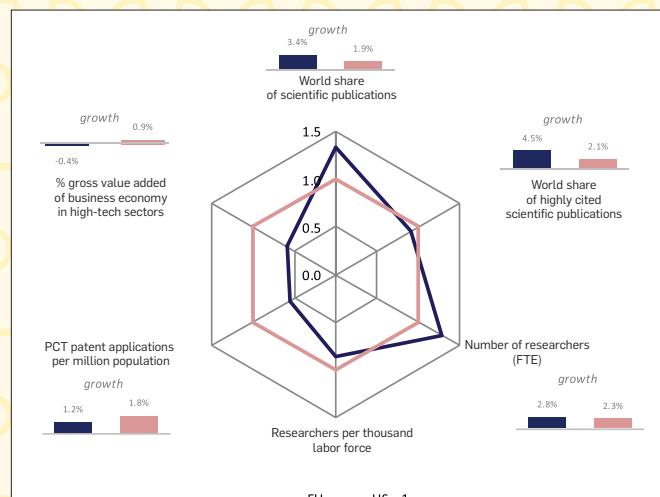
I foresee a time when learning is a lifelong commitment, with people spending 10%–20% of their work time continuing their education in order to learn new skills and make oneself relevant as the future unfolds.

zucato⁵ to target the overall goals of research to address the forefront of development. Horizon 2020 aims for challenges, while Horizon Europe seems to focus more toward missions. A possible mission agreed by many is to join forces to achieve the Sustainability Development Goals.⁶

► *Increasingly complex research problems.* The challenges on today's research agenda are becoming bigger and more complex, requiring large multidisciplinary collaboration teams. While traditional research strives to limit and focus research questions to problems that could derive scientific conclusions beyond any doubt, today's problems are becoming increasingly more "wicked,"⁷ without the opportunity to isolate particular phenomena for empirical testing. To research such problems requires methodologies and processes to develop knowledge under these conditions, ensuring scientific rigor, quality, and ethical standards being met.

► *Multidisciplinarity opportunities and challenges.* Today's complex problems require a more genuine and open collaboration across a wide range of different disciplines. It requires each individual builds a broad understanding of the context of research far beyond the disciplinary research agenda in addition to the depth required within their own field.¹ It requires true transdisciplinary work, including social science and humanities (SSH) from the outset, to be able to tackle complex problems in technical and medical research and innovation in a meaningful way.

► *Unprecedented technological development.* The last decades of development has



Comparative and growth rates of scientific publications, highly cited scientific publications, researchers, patent applications and valued-added of high-tech sectors in the EU compared to the U.S.

seen the birth and growth of many groundbreaking innovations that changed our everyday lives. Advancements have had an impact not only on the technology itself, but also on disciplines such as medicine, SSH, economy, and basic science. There is no reason to think development will slow down, rather we must embrace and support the development and maintain high ethical standards in the development.

► *Innovation.* One of the major political arguments for investing in research is that it will eventually lead to the development of new products, services, or knowledge with the potential to create new companies, employment opportunities, and eventually contribute to the economic growth of the society. The mechanisms for supporting more disruptive innovation is essential.¹⁰

► *End-user involvement and citizen science.* An essential characterizing aspect of future research and innovation is the need to incorporate and involve the general public to a much larger extent and engage in so-called citizen science.² People may

become involved as participants, for example, in more action-oriented research projects sharing their personal data.³ To be able to address our future challenges, they will become reflective practitioners⁸ in the analysis and reconstruction of the society.

► *Connection to education.* Digitalization means changing the ways we educate by providing opportunities to offer education to everybody. I foresee a time when learning is a lifelong commitment, with people spending 10%–20% of their work time continuing their education in order to learn new skills and make oneself relevant as the future unfolds. These needs must be addressed when building the research community, as there is a tight and important connection between research and education.

► *Research leadership.* There is a need for leadership that understands transdisciplinary research and knows how best to engage participants. Trustworthy and engaging leaders can guide teams through complex, wicked problem solving.

A value-driven research

process, recognizing the European values of participation, gender equality, and low corruption has been powerful throughout Europe. The key factors in reducing inequality include a strong focus on education, health, social protection, progressive taxation, higher wages for the general workforce, stronger labor rights, especially for women.⁶ These values makes Europe a unique place to develop research that features strong human and humanistic values, a strong commitment to the U.N. sustainability goals, and recognizing the opportunity for overall participation on equal terms beyond hierarchies, knowledge levels, education, or assets. □

References

- Brown, R.R., Deletic, A. and Wong, T.H.F. Interdisciplinarity: How to catalyze collaboration. *Nature* 525, 315–317 (Sept. 17, 2015) doi:10.1038/525315a
- Irwin, A. *Citizen Science: A Study of People, Expertise and Sustainable Development*. Routledge, 2002.
- Kemmis, S., and McTaggart, R. *Participatory Action Research: Communicative Action and the Public Sphere*. Sage Publications Ltd., 2005.
- Lamy, P. LAB-FAB-APP: Investing in the European Future We Want. European Commission, Luxembourg, 2017; <https://bit.ly/2sEIMKp>
- Mazzucato, M. Mission-Oriented Research and Innovation in the European Union—A problem-solving approach to fuel innovation-led growth. European Commission; https://ec.europa.eu/info/sites/info/files/mazzucato_report_2018.pdf
- Oxfam. The commitment to reducing inequality index 2018—A global ranking of governments based on what they are doing to tackle the gap between rich and poor; www.oxfam.org
- Rittel, H.W.J. and Webber, M.M. Wicked problems. *Man-made Futures* 26, 1 (1974), 272–280.
- Schön, D.A. *The Reflective Practitioner: How Professionals Think in Action*. Routledge, 2017.
- United Nations. *Transforming our World: The 2030 Agenda for Sustainable Development* (UN), 2015; <https://sustainabledevelopment.un.org>
- Von Hippel, E. *Democratizing Innovation*. MIT Press, Cambridge, MA, 2005.

Jan Gulliksen is a professor of human computer interaction and vice president for digitalization at KTH Royal Institute of Technology, Stockholm, Sweden.

Copyright held by author/owner.

HiPEAC: A European Network Built to Last

BY KOEN DE BOSSCHERE, MARC DURANTON, AND MADELEINE GRAY

HIGH PERFORMANCE AND Embedded Architecture and Compilation (HiPEAC) was founded in 2004 as a European research network. In the last 15 years, it has grown from 70 to 2,000 computing specialists, including 200 from industry, making it the largest such network in the world. Membership is free, but members are expected to be active participants in the network.

Today, HiPEAC is a hub for European researchers and industry representatives in the full range of computing systems (from sensor nodes to exascale systems).^a It has received uninterrupted funding from the European Commission for helping to implement Europe's policy to strengthen the computing community throughout the region.^b

a <https://www.hipeac.net/>
b HiPEAC is funded by the European Commission under grant agreement 779656.



The HiPEAC staff hosts a one-week summer school program for computer architects and tool builders working in the field of high-performance computer architecture and compilation for computing systems.

The HiPEAC Conference is the second largest European research gathering in computing and it is the flagship event of the network. It pioneered the innovative journal-first publication model, which means ACM's *Transactions on Architecture and*

Code Optimization (TACO) evaluates and publishes papers submitted for the HiPEAC Conference and authors of accepted papers receive an invitation to present their work at the conference. This year, all papers published in ACM TACO will be open access, which follows the requirement that all published work resulting from European-funded research programs must be open access. By combining the conference with a rich program of workshops, tutorials, poster sessions, and an industry exhibition, the HiPEAC Conference has become the premier networking event for the European comput-

ing community.^c

The ACACES summer school is another major event hosted by HiPEAC, attracting more than 200 attendees from academia and industry for a full week of advanced courses taught by world-class experts. Many attendees have credited ACACES as a life-changing event at the start of their career.

Since 2012, HiPEAC has actively invested in attracting and retaining talent in Europe. The network's careers services, in combination with its jobs and internship portal,^d support HiPEAC members in their

The HiPEAC conference has become the premier networking event for the European computing community.

search for skilled talent, or to land the perfect job in Europe. In particular, we try to match around 200 Ph.D. students yearly with the many open computing positions throughout Europe.

HiPEAC Vision, an influential biennial roadmap report produced by the community, contains a detailed SWOT-analysis of the European computing industry, including trends in the market, society, and in science and technology and provides a series of recommendations to strengthen Europe's position in the field. It is one of the key inputs in defining future research programs at the European level.^e

Lessons Learned

HiPEAC's impact in Europe's computing community takes many paths. Beyond the services it provides members, it also fosters more international collaboration, a larger supporting network, and has lessened the region's brain drain. For example: HiPEAC has distributed approximately 400 mobil-

ity grants, which have led to numerous scientific collaborations as well as the creation of start-ups and permanent hires. The HiPEAC jobs portal publishes over 500 career opportunities per year. Over 200 scientific projects have used the HiPEAC platform for promotion, of which 50 are HiPEAC stakeholder members. The network also functions as an effective bidirectional communication channel between European policy makers and the European research community at large.

It took HiPEAC about 10 years to create a strong and attractive brand, to build a large and vibrant community, and to discover the services the community valued most. HiPEAC's decision to hire a dedicated, professional staff to run the network rather than rely solely on volunteers was a turning point. Building an effective international network requires time, resources, vision, and perseverance. It simply cannot be done in a couple of years, it is difficult to accomplish with only volunteers, and

It took HiPEAC about 10 years to create a strong and attractive brand, to build a large and vibrant community, and to discover the services the community valued most.

it cannot survive without funding.

Europe is neither the U.S. nor China. It takes its own approach to building a robust, innovating computing community, one that fits the European strengths, and provides answers to European challenges and European ways of thinking. In addition to investing in technology areas like artificial intelligence, cyber security and cyber-physical systems, Europe should also invest in its small and medium-sized enterprises, which form the backbone of its economy. For example, finding solutions for societal challenges such as the rapidly ageing population

by developing the "silver economy" and investing in healthcare technologies, or focusing on its low economic growth by promoting industry 4.0; or facing environmental issues by addressing the United Nation's sustainable development goals. Europe should also invest in retraining programs for workers and in improving the digital skills of the global population. □

Koen De Bosschere is a professor at Ghent University, Belgium, and HiPEAC coordinator.

Marc Duranton is a research fellow at CEA, France, and HiPEAC vision coordinator.

Madeleine Gray is dissemination officer at Barcelona Supercomputing Center, Spain, and HiPEAC communications officer.

Copyright held by authors/owners.



A presentation session during HiPEAC 18 in Manchester, U.K.



Attendees of HiPEAC 17 congregate at the Waterfront Congress Centre in Stockholm, Sweden.

ACM Europe Council's Best Paper Awards

BY JOAQUIM JORGE, MASHHUDA GLEN CROSS, AND AARON QUIGLEY

THE ACM EUROPE Council's remit is to support European ACM members while increasing the level and global visibility of ACM activities throughout Europe. Toward this goal, the ACM Europe Council's Best Paper Awards aim to achieve three key objectives: First, is to foster, recognize, and reward research excellence showcased at ACM-sponsored conferences held in Europe. Second, is to expand awareness of the many high-quality, ACM-sponsored events that take place annually within Europe. Third, is to enhance diversity and inclusion of European research across the global ACM community of researchers, students, and practitioners.

The ACM Europe Council's Best Paper Awards recognize authors of outstanding technical contributions to ACM-sponsored conferences held in Europe. In addition, these awards



Xiao Han receives the ACM Europe Council Best Student Paper Award from Gabriele Anderst-Kotsis during CCS 2016 in Vienna, Austria.

acknowledge ground-breaking research in each conference's discipline for its importance and contribution to computing, and to highlight theoretical and practical innovations likely to shape the future of computing both within Europe and globally. This initiative began in 2016 as

an award to recognize best student papers. In 2018, the ACM Awards Committee recognized the initiative as a meritorious endeavor that now extends to both junior and senior participation categories. Thus far, eight awards have been bestowed, and six more are planned for the fiscal year 2019 including ACM CHI19 in Glasgow. Currently, this is the only regional-based best paper award of its kind offered by ACM.

There are between 30 and 50 ACM-sponsored conferences held annually in Europe. These events bring together a community of 6,000–10,000 researchers, students, and professionals. While every ACM-sponsored conference

that takes place in Europe is eligible for the award, to date only a few have been invited to confer this distinction. Among those, for example, is the ACM Conference on Computer and Communications Security (CCS, the flagship annual conference of ACM's Special Interest Group on Security, Audit and Control or SIGSAC), which bestowed this award in 2016. From its inception, CCS has established itself as a high standard research conference in its area. The ACM Europe Council was proud to join SIGSAC in awarding the Best Student Paper to Xiao Han, Nizar Kheir (Orange Labs), and Davide Balzarotti (Eurecom) for their work "PhishEye: Live

The ACM Europe Council's Best Paper Awards recognize authors of outstanding technical contributions to ACM-sponsored conferences held in Europe.

Monitoring of Sandboxed Phishing Kits.”^a

Another well-established venue is ACM Virtual Reality Software Technology, one of the oldest conferences in the field of virtual reality. The ACM Europe Council was proud to join SIGCHI/SIGGRAPH in bestowing the Best Paper Award to Misha Sra, a student from MIT’s Media Lab Student, Sergio Garrido-Jurado, from the University of Córdoba, Chris Schmandt, and Pattie Maes from MIT’s Media Lab for their paper entitled, “Procedurally Generated Virtual Reality from 3D Reconstructed Physical Space.”^b This award aims to foster excellence, and as recipients

a <https://dl.acm.org/citation.cfm?id=2978330>

b <https://dl.acm.org/citation.cfm?id=2993372>

Garrido-Jurado and Sra commented: “The award was a validation of an idea we both strongly believed in despite some people telling us otherwise.”

The ACM Europe Council was also honored to bestow on researcher Sergio Cabello, his first Best Paper Award in recognition of his work, “Subquadratic Algorithms for the Diameter and the Sum of Pairwise Distances in Planar Graphs,” at the ACM-SIAM 2017 Symposium on Discrete Algorithms. This distinction aims to expand awareness of ACM conferences held in Europe, and as Cabello noted, “I definitely got a lot of recognition from colleagues in my research area because of the award.”

For most of its recipients, the ACM Europe Council’s

By showcasing such work, and underscoring the diversity of research areas represented at ACM-sponsored conferences held across Europe, this award highlights excellence to the global ACM community.

Best Paper Award was their first such distinction. It will be interesting to see what lasting impact this recognition will have on the awardees. From their initial feedback, the short-term personal impacts have ranged from “more recognition from colleagues,” “more speaker invitations,” “more requests to review

papers,” “wider recognition for authors within their institutions and internationally,” “confidence to continue to work on projects,” “a first step of a European funded project,” “follow-on journal papers,” and “encouragement and validation.” By showcasing such work, and underscoring the diversity of research areas represented at ACM-sponsored conferences held across Europe, this award will continue to highlight excellence to the global ACM community.

For more details about the award, and to view distinguished papers to date, visit <https://europe.acm.org/awards>. □

Joaquim Jorge is a professor of computer graphics and multimedia at the Departamento de Engenharia Informática do Instituto Superior Técnico da Universidade de Lisboa in Lisbon, Portugal. He is also secretary of ACM Europe Council.

Mashhuda Glencross is the director for research and development at Switch That Technologies Ltd. and Pismo Software Ltd., both in Oxford, U.K. She is also Director-at-Large for ACM SIGGRAPH.

Aaron Quigley is chair of Human Computer Interactions at the University of St. Andrews, Scotland. He is SIGCHI vice president for conferences and general co-chair for ACM CHI 2021.

ACM Europe Council seeks to continue its engagement with ACM conferences held in Europe. If you are chairing such a conference and would like to be considered for the ACM Europe Council’s Best Paper Award, please contact jorgej@acm.org.

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



Misha Sra receives the ACM Europe Council Best Student Paper Award from Hans-Joachim Hof during VRST 2016 in Munich, Germany.

BY JULIE A. MCCANN, GIAN PIETRO PICCO,
ALEX GLUHAK, KARL HENRIK JOHANSSON,
MARTIN TÖRNGREN, AND LAILA GIDE

Connected Things Connecting Europe

IT IS ESTIMATED that personal computers, datacenters, and other technologies constitute less than 1% of all microprocessor usage;¹⁰ embedded systems represent the remaining percentage and can be found in our washing machines, microwaves, remote controls, and PC peripherals (such as keyboards and mobile phones), with modern cars containing many tens of embedded microcontrollers.¹⁸ Modern embedded-system microcontroller and transceiver technology advancements have brought forth the kinds of systems we have in the past defined as pervasive, ubiquitous, and embedded computing, and for some time in Europe, “embedded intelligence.” However, today they are better known as the Internet of Things (IoT) and cyber-physical systems (CPS); see the figure here.

The jury is still out regarding a definition of the latter two terms or indeed how to differentiate them, but people generally tend to refer to IoT as embedded devices that connect to the Internet



to exchange data, optimize processes, monitor environments, and typically consist of sensors, actuators, and low-power compute infrastructures. CPS is a term first coined in 2006 in the U.S. to characterize “the integration of physical systems and processes with networked computing” for systems that “use computations and communication deeply embedded in and interacting with physical processes to add new capabilities to physical systems.”²² CPS is generally put forward as the more systems notion, while IoT emphasizes communication and analytics, yet IoT-like devices need not always use Internet protocols to create a CPS, hence the ambiguity. The European Commission debated for two years whether to call its embedded intelligence programs CPS, with the latter winning out in the end. In this article, we embrace these terms fluidly and name them IoT/CPS; for other definitions, see the sidebar “Some Definitions.”

In essence, the terms represent different perspectives on the technological advancements that have led to creation of many related application terms—industry 4.0, smart cities, preci-



sion agriculture, smart transport, and autonomous vehicles—all representing new classes of technologically enabled systems. Recent studies have predicted the impact of IoT/CPS on the European Union's GDP in 2025 by sector, with “transportation” being forecast to create the greatest value, with a total of €245 billion alone, followed closely by “healthcare,” “housing,” and “industry.”¹ As in the rest of the world, European countries and the European Commission have invested heavily in IoT/CPS research, almost €200 million, resulting in many cross-discipline, cross-country technology advancements unique in terms of their focus on the integration of such systems, particularly at scale, their underpinning communications substrates, and, more recently, their security and relationship to privacy. In this article we describe highlights of this work in more detail and present what we believe are the main outstanding challenges facing the field for Europe over the next decade.

The Integrated Approach

Two decades ago, a European named

Kevin Ashton coined the phrase the “Internet of Things.” His vision of connecting sensor- and actuator-based technology to the Internet unfolded an active area of technological advancement around the world that only in the past few years has begun to find larger-scale adoption and is finally becoming a commercial reality. In parallel, the University of California, Berkeley’s TinyOS and Mote hardware combination dominated early academic experimental work on wireless sensor networks, also making its way into various commercial products. While the U.S. focused on designing new protocols and approaches to overcome the intrinsic limitations of resource-constraint IoT/CPS devices, the focus in Europe was more on the integration of such systems to fulfil real application needs. In particular, tools to aid the building of devices and, moreover, their applications became the European emphasis, resulting in mechanisms to ease programming and systems engineering and, more important, make such systems a natural extension of the Internet, the latter

reaping the advantages of standardization and the software-engineering experience that programmers already had from other Internet systems.

Adam Dunkels of SICS in Sweden developed the Contiki operating system that has significantly grown in popularity, particularly over the past decade. As the technology matured, device hardware could pack more compute power for the same energy budget, and the resource savings TinyOS delivered thus became less important. Exploiting this power, Contiki’s advantage over its predecessors lay in its flexibility and ease of coding applications. Indeed, today hackers, academic institutions, and companies are using Contiki because it remains lightweight, mature, and free; for example, Texas Instruments (a U.S. company) ships many of its IoT/CPS devices (such as Sensortag) with the option of using Contiki.

At that time, the prevalent communications protocols operated over low-data-rate, local-area networking (up to approximately 50-meter distances) radio transmissions. 802.15.4-based pro-

tocols (such as ZigBee, designed in the U.S.) overcome these short distances by providing multi-hop communications, allowing data to be relayed between devices to form longer-distance routes and hopping the data from device to device. However, in Europe, for researchers (such as Dunkels), the quest was now to push the Internet Protocol all the way down to small embedded devices themselves.² Early attempts include the work of Zach Shelby, an American, working at Oulu University,³ but the now-ubiquitous 6LowPAN protocol has emerged to provide lightweight end-to-end Internet connectivity down to the smallest devices and has gradually replaced the previously popular ZigBee communications approach.

While 6LowPAN allowed for more efficient raw Internet communication, the next logical step was to make it Web friendly by replacing its heavy Web protocol with a more lightweight one. Work emanating from European large-scale mixed academic/industrial projects (such as SENSEI and FP7) focused on how such emerging wireless sensor and actuator networks can be more effectively integrated into a future Internet.⁴ Shelby worked with the Internet Engineering Task Force (IETF) and such companies as England's ARM low-power processor design company, ultimately producing the COAP protocol⁵ used to make applications on low-power devices easier to program. At about the same time, Dom Guinard at

ETH Zurich and others advocated for such devices to become first-class citizens in the current Web. His pioneering work led to what is now known as the Web of Things, with active work (as part of W3C) receiving support from Siemens, Google, and other sources.⁶

The increasing investment in CPS/IoT throughout Europe, and the world, has meant an increase in the number of systems, protocols, and applications being built. Also, there was little integration between systems, as seen especially in the smart-city domain. This fragmentation is thought to have undermined the confidence of stakeholders and market opportunities, affecting IoT adoption, thus causing Europe to become increasingly focused on the applications built on top of IoT/CPS systems and their integration.⁷ Indeed the perception in Europe was that while U.S. IoT/CPS innovation focused on adoption environments based on individual business cases driven by economic return on investment, Europe's industry and academic researchers focused on the exploration of societal benefits and acceptance of CPS/IoT technology generally.

One major success resulting from the European joint research and industrial projects is FIWARE, a curated framework of open source-platform, market-ready components to accelerate development of IoT/CPS systems and their integration with cloud services.⁸ Since its beginnings in 2012, FIWARE has evolved from a consor-

tium of multinational telcos, including Telefónica (Spain), Orange (formally French Telecom), and others, to a suite of more than 50 components to create value from real-world applications enabled by the ubiquity of heterogeneous and resource-constrained devices. Another example is from the ARTEMIS Industry Association,⁹ with more than 170 members and associates from all over Europe, and the European IoT Platform Initiative Programme (IoT EPI), a €50 million programme with nine projects involving more than 40 different IoT platforms exploring multiple approaches to interoperability.¹¹

An example of where Europe leads in living-lab deployments is the SmartSantander testbed in Santander, Spain, a prominent European experimental infrastructure for IoT/CPS. By embedding a large number of diverse sensor devices into a city environment, it allowed a variety of smart city use cases to be explored. While initially useful for experiments with IoT protocols and data-driven services, the infrastructure is now part of the Santander's day-to-day operation, improving the lives of its citizens. Since its beginnings in 2010, more than 12,000 sensor devices have been deployed across the city to help the government operate as efficiently as possible through such applications as adaptive traffic management, smart parking, water management, intelligent streetlights, and waste disposal. SmartSantander went on to inspire other initiatives around the world, in-

Timeline of some key IoT events.



cluding the Array of Things project in Chicago.

Both testbeds and living labs²¹ paved the way for IoT large-scale pilots in Europe, a €100M R&I program that commenced in 2017.¹² Examples of such projects include SynchroniCity (eight smart-city pilots¹³), MONICA (IoT technologies to manage sound and security at large, open-air cultural and sporting events¹⁴), and IoF2020 (Internet of Food and Farm 2020 with 70 partners from 16 European countries¹⁵); many of these projects are associated with, and continue to use, the FIWARE infrastructures.

Underpinning Communications Technologies

The communications substrate in an IoT/CPS architecture plays a crucial role, and low-power wireless connectivity is fundamental to balancing connectivity performance with low-power system capabilities and lifetimes. Freedom from power and data cables provides mobility and autonomy of devices that are readily deployed and relocated, can improve performance, and follow the users and objects they are attached to, or even move of their own volition, as with robots and drones. In recent decades, wireless communications was dominated by Wi-Fi and cellular communications that were ubiquitous yet energy-hungry; low-power alternatives had to emerge, as embodied by ZigBee in 2004. Today, the wireless communication landscape is significantly more

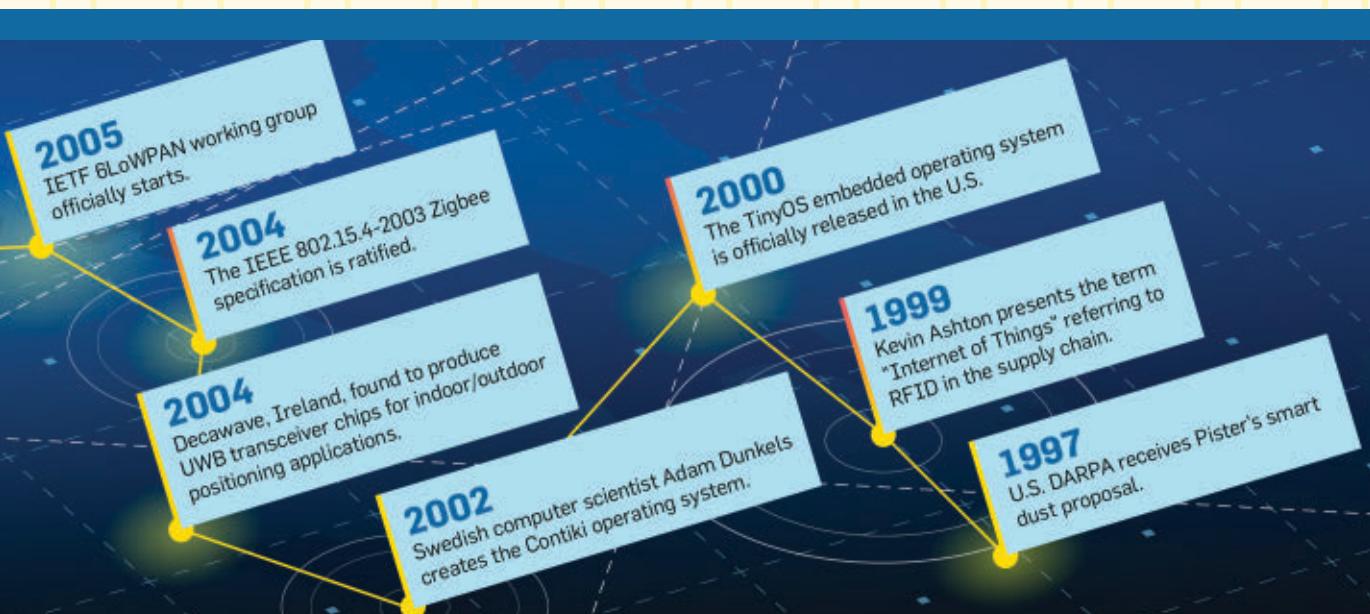
fluid, with several technologies (both competing and complementary) offering disruptive opportunities unthinkable only a few years ago. Interestingly, several of these trends are the result of achievements by European researchers and companies, as we highlight here.

A prominent example is a new class of communications mechanisms described as “low-power wide-area networks” (LPWANs) that recently revealed trade-offs in the amount of power they require from the device, geographic coverage or distances they send data, and data rates. Until a few years ago, long-range communication was a privilege of cellular telephone communication, where devices were fitted with SIM cards and communicated typically over 2G GPRS networks. This did not match with the low-power nature of CPS/IoT devices and meant they were required to be plugged into the mains, limiting where they could be placed or receive frequent battery changes, and many stakeholders were reluctant to rely solely on proprietary networks and devices owned by operators.

SigFox¹⁶ based in France was in 2009 the first to use ultra-narrowband modulation to enable longer-distance communications while remaining low power. Since the first deployments that covered the entire country of France, SigFox showed its technology can provide coverage like cellphone communications but without the need for a SIM card and at significantly less cost in terms of money and energy. But

SigFox is still a telco operator, having to manage access to its own network and based on proprietary technology. In contrast, LoRa,¹⁷ which was developed by Cycleo of Grenoble, France, and acquired by Semtech in 2012, used radio technology based on chirp spread spectrum modulation to effect low-power wide-area transmission. The LoRa Alliance then defined a public suite of protocol specifications (LoRaWAN) that allows a telco operator to deploy its own networks but also enables deployment and operation of privately owned networks operating side-by-side. Both SigFox and LoRa have their main center of gravity in Europe; for instance, of the 5,000+ gateways deployed today, 3,000+ are in Europe. This is also reflected in the surge of competing, industry-driven approaches, among which, arguably the most prominent, is Huawei’s NB-IoT. Indeed, today’s version of NB-IoT, which is being specified by the 3GPP, an international body of telcos, originated in early work by NEUL, a company from Cambridge, U.K., that developed the Weightless protocol and was bought out in 2014 by Huawei. LPWA technologies are not being rolled out worldwide.

Where LPWA supports slow data over great distances, ultra-wideband (UWB) communications permits higher data volumes and speeds over short distances. Originally used for military applications, UWB became unlicensed in 2002, but a new wave of interest has followed a small Irish company called DecaWave¹⁸ when it released the



Some Definitions

Microcontroller. Computer on a single chip, with one or more processor cores, memory, and input/output peripherals.

Sensor nodes/mode. Generic way to describe sensor-based devices, typically consisting of several sensors and radio communications module(s) governed by a microcontroller. Different from phones and traditional computers, they are a few centimeters in size without keyboard or screen. An example is the University of California, Berkeley, TMote Sky sensor node consisting of the CC2420 ZigBee near-range communications, an MSP430 low-power microcontroller packed into a matchbox-size form factor.

Actuator. A device that controls other devices (such as valves and switches).

European Research Council. A body that funds technological research in the EU. Its framework funding programs include FP7 (Framework Programme) finished in 2013, giving way to H2020 (Horizon 2020). On top of this, each EU country also has national funding infrastructures, as in EPSRC in the U.K. and DFG in Germany.

IETF. The IETF is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of Internet architecture and smooth operation of the Internet; for more, see <https://www.ietf.org/about/>

DW1000 chip, overcoming many of bulkiness and power-consumption issues and storming the field of real-time location-tracking systems. Indeed, the potential here is enormous, especially if UWB chips eventually find their way into smartphones where UWB could trigger a new wave of location-based IoT/CPS services with an impact comparable to (if not greater than) that achieved by GPS.

Trust, Safety, Security, Privacy (Guarantees)

CPS and IoT provide unprecedented capabilities and opportunities for the benefit of society. But it will be achieved through corresponding unprecedented technological complexity that also introduces new risks that need to be recognized, debated, and dealt with appropriately. This is essential since future CPS and IoT will be widespread and underpin a large number of critical societal infrastructures, including water, energy, transportation, and healthcare, all relying on the proper operation of the technologies.

A key concern is that current engineering methodologies are generally viewed as inadequate for next-generation CPS. Consequently, multiple calls have been issued from the EU for new methodologies, including Platform-4CPs,²⁵ AENEAS,²⁶ and the Acatech National Academy of Science and Engineering.²⁷ The full potential of future CPS can be obtained only when new engineering methodologies are in place

to ensure future CPS systems are sufficiently safe, secure, available, privacy-preserving, and overall trustworthy. A science for CPS engineering is needed. Europe is positioned well in this regard to address the key challenges of complexity management, safety, and security by design and privacy.

Complexity management of IoT/CPS systems is important because they inherit the complexity of their cyber and physical parts. There is a lack of approaches to systematically accomplish “composability” of CPS components, meaning achieving integration of CPS components is difficult without negative, sometimes unknown, side effects, or emerging behaviors.²⁸ Composability for CPS must address the multifaceted dependencies in CPS across components, functions, and system-level properties. An example of a European stronghold is the effort driven by Kopetz on composable time-triggered architectures, with research funded through several EU projects that have influenced many communication protocols for CPS, delivered reusable architectural services for exploitation across platforms of different domains (INDEXYS project in 2008), and paved the way for successful companies like TTTech.²⁹

The use of machine learning, particularly deep learning, provides a novel technology within CPS. While such technologies enable entirely new types of applications, they raise concerns about how to deal with transparency

(black-box behavior), robustness, predictability (such as when data is outside a training set), and how to cost-efficiently verify, validate, and assure such systems.^{29,30} In addition, CPS systems must function in increasingly complex environments, as in automated driving. Describing such varying environments and systematically dealing with uncertainty represent further key challenges that have been addressed in such European research projects as Pegasus³¹ and the U.K. EPSRC-funded S4: Science for Sensor Systems in 2016.³²

Safety and security engineering concerns the connectivity and spread of CPS and provides new attack surfaces that could exploit vulnerabilities in the cyber and/or physical side, as well as among human stakeholders. This implies existing security approaches are not suitable. Moreover, security may affect safety, thus calling for integrated and balanced security and safety trade-offs and development of new methodologies. The widespread use of CPS systems and their increasing automation imply that existing safety-engineering approaches are not sufficient, and, in particular, that future CPS will need to deal with risk explicitly, incorporating measures of dynamic risk, as compared, again, with automated driving. An example of security research in Europe comes from the £23M PETRAS Research Hub in the U.K., which involves 60 projects researching the various aspects of IoT/CPS security, from devices to social practice, and have produced a landmark report, *The Internet of Things: Realising the Potential of a Trusted Smart World*,³³ co-produced with the Royal Academy of Engineering.

It is infeasible to predict all possible faults, threats, and failure modes for future CPS. Systems will have to be resilient, with built-in build monitors and error handlers to ensure cost-efficient dependability. Examples of European efforts include the MBAT project that gave European industry a leading-edge affordable and effective validation-and-verification technology in the form of a Reference Technology Platform (the MBAT RTP) and the AQUAS project, which is developing solutions for safety/security/performance co-engineering, as in Sillitto.³⁴ Europe has a strong tradition

tion in dependability and engineering of trustworthy systems, notably through the ARTEMIS and ECSEL private-public partnerships. Example projects include Pegasus, funded by the German Federal Ministry for Economic Affairs and Energy and involving all major German OEMs and Tier 1 companies to produce mechanisms to test and formally verify autonomous vehicles. And SCOTT is examining frameworks to enable development of secure and connected trustworthy things primarily for the rail-transport industries.³⁵ Separately, the TrustLite security framework from the Intel Collaborative Research Institute for Secure Computing (a collaboration of TU Darmstadt, University of Helsinki, and other European security institutes) have produced an Execution-Aware Memory Protection Unit that provides programmable operating system-independent isolation of software modules at runtime for low-cost embedded devices.

IoT/CPS systems are constantly monitoring homes, factories, cars, and more, and while understanding these processes can make them more efficient, sustainable, and safe, they can expose privacy concerns. The most prominent European initiative affecting IoT/CPS data gathering is that of the General Data Protection Regulation (GDPR) regarding data protection for individuals in the EU and its economic area.³⁶ The European approach to privacy is that, through GDPR, all the requirements of data domains and territories are consolidated into a single coherent and well-defined regulation. One aspect of this is that a data owner must prove its data protection reasonably matches the current state of the art, which in turn uniquely drives practical anonymization research. Researchers aim to demonstrate privacy shortfalls to make schemes more robust. For example, U.K. and Belgium researchers³⁷ were able to prove it took only four location points to be able to uniquely identify someone 95% of the time and that data coarsening and noise addition do not help. This was followed by Gadotti et al.,²³ who showed privacy techniques using “sticky noise” could be easily infiltrated. All European citizens, as well as those only visiting Europe, are covered by GDPR, meaning its

effect reaches much farther than just Europe.

Conclusion

We have drawn out three views of IoT/CPS systems the European approach to research contributes to in its own unique way, though European researchers continue to collaborate across the globe to address the many challenges associated with these systems. This subject continues to grow and, with it, new problems. For example, as such systems contribute to the autonomy of cars, water networks, precision farms, and more, the more we need to be able to understand how to engineer them and provide guarantees regarding their operation. However, as we do not fully understand how digital systems interact with the physical world, we do not yet have such guarantees. We thus need a science of cyber-physical interaction; related design principles will then emerge, much as they have in other engineering disciplines. Given the importance of the communications substructure for such systems, the jury is still out as to which protocol (or set) will win.

There are many players in the LPWA game, but the big question is what will be the effect of the promised 5G suite of solutions? Finally, as these systems take more control of our lives, their ethical approach is key, including the ability to maintain privacy while still being useful. Indeed, their security is of paramount importance, as being able to hack a water network or autonomous vehicle could mean disaster. There is plenty of research for Europe and beyond to consider. C

References

- <https://www.statista.com/statistics/686173/iot-s-impact-on-gdp-in-the-european-union-eu-by-sector/>
- Dunkels, D. Full TCP/IP for 8-bit architectures. In *Proceedings of the 1st ACM/Usenix International Conference on Mobile Systems, Applications and Services* (San Francisco, May 2003).
- Shelby, Z., Mahonen, P., Riihijarvi, J., Raivio, O., and Huusko, P. NanoIP: The Zen of embedded networking. In *Proceedings of the IEEE International Conference on Communications* (Anchorage, AK, 2003), 1218–1222.
- <https://tools.ietf.org/id/draft-shelby-core-coap-req-01.html>
- <http://coap.technology/>
- http://www.usa.siemens.com/en/about_us/research/web-of-things.htm and <https://github.com/google/physical-web>
- http://www.internet-of-things-research.eu/pdf/IERC_Position_Paper_IoT_Semantic_Interoperability_Final.pdf
- <https://www.fiware.org/about-us/>
- <https://artemis-ia.eu/project/59-3car.html>
- Turley, J. Embedded processors by the numbers.
- EE Times* (May 1, 1999); https://www.eetimes.com/author.asp?section_id=36&doc_id=1287712
- <https://iot-epi.eu/>
- <https://european-iot-pilots.eu/>
- <https://synchronicity-iot.eu/>
- <https://european-iot-pilots.eu/project/monica/>
- <https://european-iot-pilots.eu/project/iot2020-2/>
- <https://www.sigfox.com>
- <https://lora-alliance.org/>
- Fleming, B. Microcontroller units in automobiles. *IEEE Vehicular Technology Magazine* 6, 3 (2011), 4–8.
- <https://www.decawave.com/>
- <http://www.indexys.eu/>
- [Open Living Labs; https://enoll.org/](https://enoll.org/)
- Lee, E.A. and Seshia, S.A. *Introduction to Embedded Systems: A Cyber-Physical Systems Approach*. MIT Press, Cambridge, MA, 2016.
- Gadotti, A., Houssiau, F., Rocher, L., and de Montjoye, Y.A. When the signal is in the noise: The limits of Diffix's sticky noise. 2018; *arXiv preprint arXiv:1804.06752*
- Istomin, T. et al. Data prediction + synchronous transmissions = ultra-low-power wireless sensor networks. In *Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems*, 2016.
- Platforms4CPS: Final recommendations; <https://bit.ly/2BodgrP>
- AENEAS, ARTEMIS Industry Association, EPoSS. *Strategic Research Agenda for Electronic Components and Systems*, 2018; <https://efecs.eu/publication/download/ecs-sra-2018.pdf>
- Acatech National Academy of Science and Engineering. *Living in a Networked World. Integrated Research Agenda Cyber-Physical Systems*, 2015; http://www.cyphers.eu/sites/default/files/acatech_STUDIE_agendaCPS_eng_ANSICHT.pdf
- Törngren, M. and Grogan, P.T. How to deal with the complexity of future cyber-physical systems? *Journal of Designs* 2, 4, 2018; <http://www.mdpi.com/2411-9660/2/4/40>
- Wagner, M. and Koopman, P. *A Philosophy for Developing Trust in Self-driving cars*. In *Road Vehicle Automation*, G. Meyer and S. Beiker, Eds. Lecture Notes in Mobility. Springer, Cham, Switzerland, 2015, 163–171.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. *Concrete Problems in AI Safety*. 2016; *arXiv:1606.06565*
- <https://www.pegasusprojekt.de/en/about-PEGASUS>
- Calder M., Dobson, S., Fisher, M., and McCann, J. *Making Sense of the World: Models for Reliable Sensor-Driven Systems*, Mar. 28, 2018; *arXiv preprint arXiv:1803.10478*
- <http://www.oerc.ox.ac.uk/news/Centre-contribution-IoT-reports>
- Sillito, H. *Architecting Systems: Concepts, Principles and Practice. Volume 6: Systems*. College Publications, London, U.K., 2014.
- <https://www.indracompany.com/en/indra/scott-secure-connected-trustable-things>
- <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- De Montjoye, Y.-A. et al. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports* 3 (2013), 1376; <http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html>

Julie A. McCann is a professor of computer systems in the Department of Computing at Imperial College, London, U.K.

Gian Pietro Picco is a professor in the Department of Information Engineering and Computer Science at the University of Trento, Italy.

Alex Gluhak is head of technology at Digital Catapult, Guildford, U.K.

Karl Henrik Johansson is a professor in the School of Electrical Engineering and Computer Science at KTH Royal Institute of Technology, Stockholm, Sweden.

Martin Törngren is a professor of embedded control systems in the Department of Machine Design at KTH Royal Institute of Technology, Stockholm, Sweden.

Laila Gide is Past-President, ARTEMIS Industry Association, and Past-Director, Advanced Studies Europe in the Corporate Strategy, Marketing and Technical Directorate, Amsterdam, The Netherlands.

Copyright held by authors/owners.

DOI:10.1145/3312565

BY PANAGIOTA FATOUROU, YOTA PAPAGEORGIOU,
AND VASILIKI PETOUSI

Women Are Needed in STEM: European Policies and Incentives

WOMEN'S PERSISTENT UNDERREPRESENTATION in science, technology, engineering, and mathematics (STEM) education, occupations, and careers in various parts of the world and its negative impact on STEM labor force and research and innovation (R&I) have given rise to measures, projects, and initiatives aimed at promoting gender equality (GE). In Europe, gender balance in R&I is understood as a social justice and equality issue. Various measures (for example, regulation and research framework, bodies, agencies, funding schemes, prizes, and awards) have been implemented at the European Commission (EC) and European Union (EU) levels to increase women's participation and include the gender dimension in R&I

Achieving GE will significantly advance the STEM labor force, research and innovation, enhance the economy, and reduce the risk of women's social exclusion to the benefit of society.

This article considers the main issues regarding GE in STEM in Europe including an analysis of the reasons for its necessity; a description of the European Union's strategy, measures, initiatives, and activities toward achieving GE; and, finally, their anticipated impact.

Gender Balance in STEM and the Necessity for Gender Equality

The recent momentous growth of the digital production sector offers extended employment opportunities for STEM and ICT-skilled employees. Within the European Union (EU), employment of STEM-skilled personnel increased by 12% between 2000 and 2013.^{4,7} According to the *Tech Nation* journal, in the U.K. alone, 1.46 million people (7.5% of the country's workforce) are employed in this sector.¹³ Future increases are anticipated. The European Commission (EC) for example, estimates that by 2020 over 900,000 additional employees will be needed in the IT sector whereas for the entire STEM sector, seven million job openings are forecast by 2025.⁴

Despite the good present employment opportunities and the future occupational prospects, European countries face a conspicuous labor shortage in STEM, which tends to be more pronounced in the digital sector.⁴⁻⁷ The number of young persons pursuing STEM-related studies is decreasing, contrary to the increasing number of university graduates, while a significant proportion of current STEM employees are approaching retirement age.⁷ Consequently, ICT- and STEM-related professions are among the top five occupations facing skill shortage in Europe. Excepting Finland, all EU member states lack such professionals.⁴ Thus, labor force availability and recruitment in the field are becoming increasingly more challenging. A con-

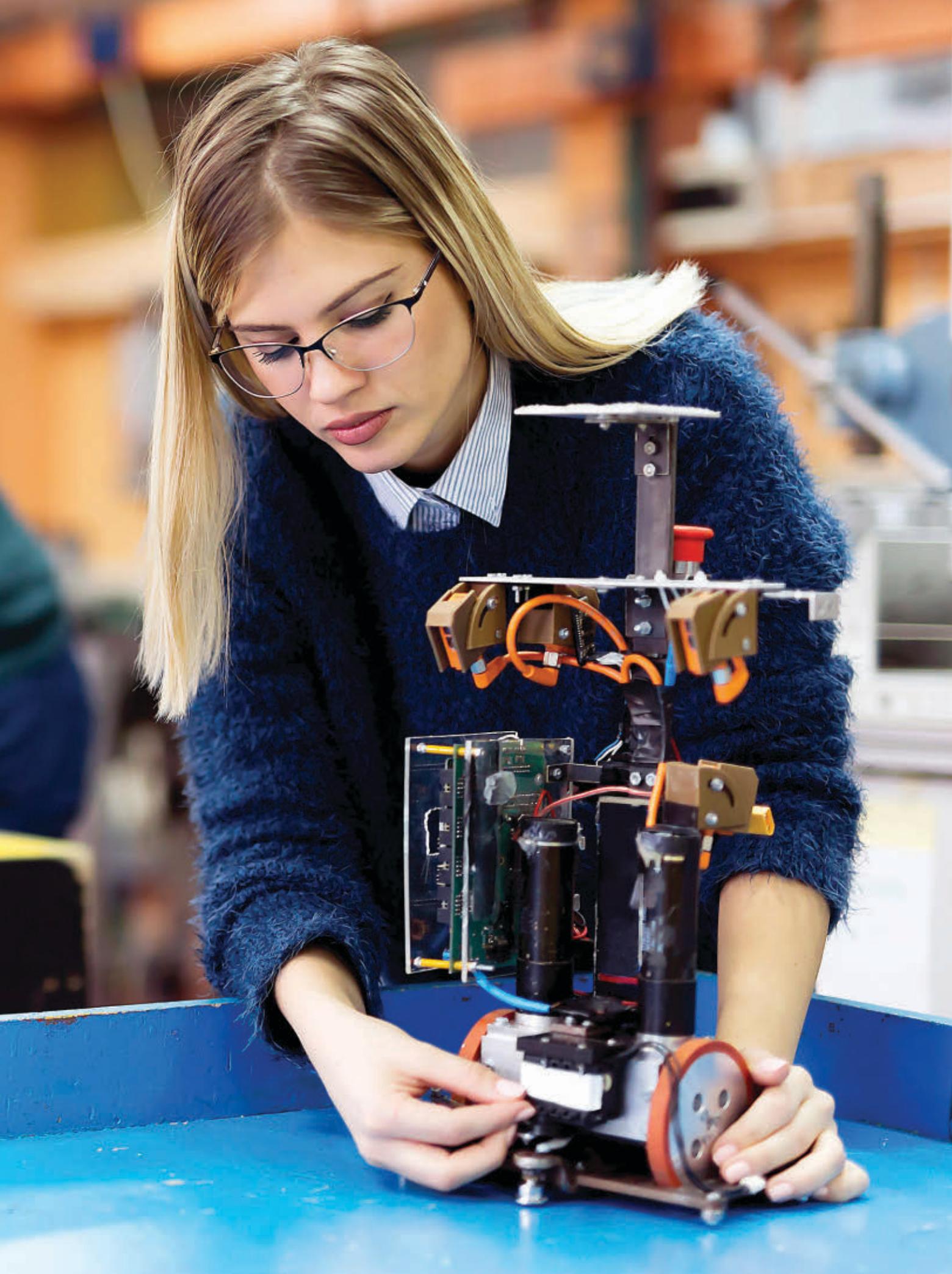


Figure 1. Informatics education in Europe: Institutions, degrees, students, positions, salaries, key data 2012–2017, October 2018.
Source: Informatics Europe.

Table 2bis graph: women percentage of students in Informatics Bachelor's programs (first year)

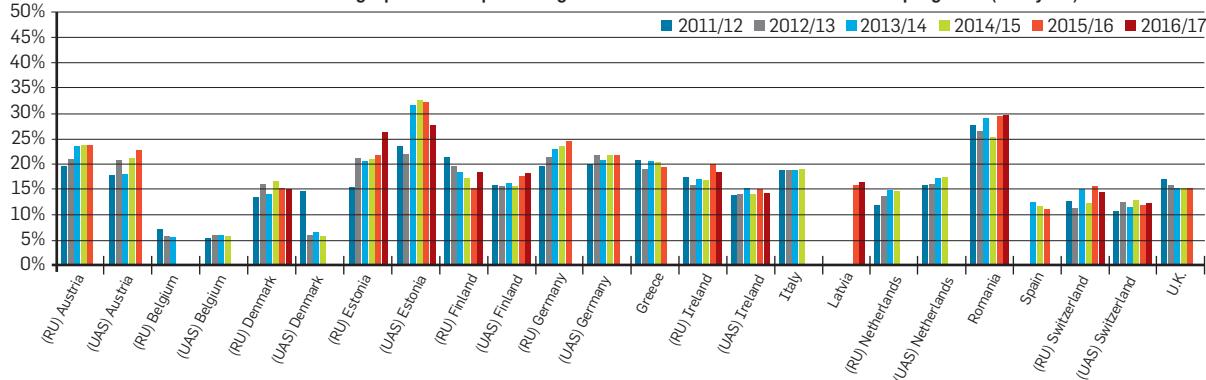
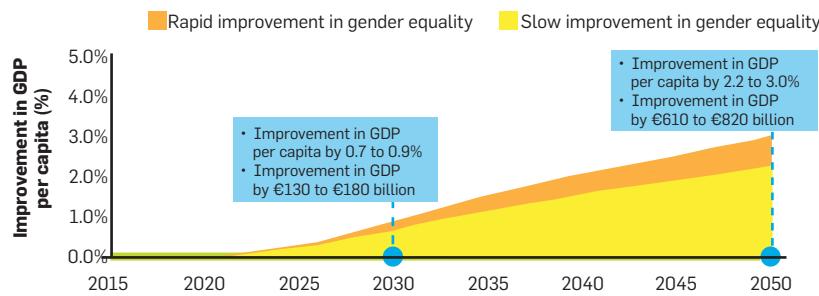


Figure 2. The effect of closing the gender gap in STEM on GDP per capita.
Source: European Commission.⁶



Indicative list of non-EC related bodies, agencies, and associations promoting gender equality in ICT in Europe.

Association	URL
ACM-Women Europe	https://europe.acm.org/
Women in Research and Education (WIRE)	http://www.informatics-europe.org/working-groups/women-in-icst-research-and-education.html
Women in High Performance Computing	https://womeninhpcc.org/
Athena SWAN	https://www.ecu.ac.uk/equality-charters/athena-swan/
Codess	https://www.codess.net/
Women in Technology and Science	https://witsireland.com/
Women in Games	http://www.womeningames.org/
European Network of Women in Leadership	https://www.wileurope.org/
European Network of Women Web Entrepreneurs Hubs	http://wehubs.eu/
Startup Europe Leaders Club	http://portal.opendiscoveryspace.eu/et/node/822017
European Platform of Women Scientists	https://epws.org/

tributing factor to this challenge is women's persistent underrepresentation.

Indeed, women who choose careers in ICT account for less than 2% of all women in the European labor market while their participation decreases with age (see <https://bit.ly/2K4Imdv>). Similarly disheartening is women's involvement in innovation and entrepreneurship. For example, women in the EU, parallel to the situation in other parts of the world, constitute less than 25% of science and engineering professionals³ and only 14% of associate professionals, that is, those who perform research and operational tasks including supervision and control of technical and operational aspects of engineering operations (see <https://bit.ly/2GqWBQI> for details on these statistics).

According to Catalyst, in 2014 women accounted for less than 1/3 of all employees in scientific research and development across the world (averaged across regions, see <https://bit.ly/2zIUJpA>). In Australia, for instance, women engineers represent less than 13% of the labor force. In Japan, despite recent measures intended to improve gender ratios in STEM, neither the 20% target of women in science, nor the 15% target of women engineers had been met by 2016. Recent reports estimate that women comprise 39.8% of all researchers in China¹² and refer to the phenomenon of the "missing women in STEM."¹⁴ Similarly, in the U.S., women earning engineering, computer, and information sci-

ences degrees represent less than 20% (data for 2014–2015) of all graduates in these fields and less than 43% in all other STEM fields (see <https://bit.ly/2zIUJpA>). In India, although gender balance has been achieved with respect to graduation rates in science, IT, and computers (data for 2015–2016), women represent less than 32% of all graduates in the engineering and technology fields. In many European countries women account for less than 20% of all students enrolled in informatics studies (see Figure 1).⁹ Concurrently, no significant progress has been observed for the past six years.

This tenacious underrepresentation of women in STEM is further manifested as gender segregation in research and science, gender-related career challenges, gender disproportions in senior positions in academia, gender imbalance in access to research funding, gender-blind and gender-biased research, and organizational culture and institutional process. As detailed in the 2012 report^a of the EC Expert Group on Structural Change, gender inequalities in research institutions are shaped by: opaqueness in decision making; institutional practices based on unconscious biases in assessment of merit, leadership suitability, and performance evaluation; unconscious gender biases in assessment of excellence and the process of peer review; gender biases in the content of science itself; and a gendered labor organization with implications for research institutions as well. It is, however, progressively acknowledged that including women in STEM studies and professions will enlarge the relative pool of skills, talents, and resources; will enhance the research process and research outcomes; will increase innovation potential; and will boost major sectors of the economy. After all, gender equality refers to equal rights, responsibilities, and opportunities for women and men and girls and boys, and entails consideration of “the interests, needs, and priorities of both women and men.”^b

Within the EU, gender equality in all

aspects of social, political, and cultural life, including education and R&I, is approached as a matter of social justice and fairness. GE is included in the EC's priorities and is defined as “promoting equal economic independence for women and men, closing the gender gap, advancing gender balance in decision-making, ending gender-based violence, and promoting gender equality beyond the EU.”^c The official policy for achieving gender equality endorsed by the EU is gender mainstreaming,^d an internationally embraced strategy that “involves the integration of a gender perspective into the preparation, design, implementation, monitoring, and evaluation of policies, regulatory measures, and spending programs, with a view to promoting equality between women and men and combating discrimination.”^e Toward this end, targeted measures have been developed and actions undertaken at the national and European levels. Although their results vary, and their full potential has not yet been realized, such measures attest to Europe's commitment to gender equality. Here, we present these efforts as they pertain specifically to STEM research and education.

Achieving gender equality will significantly advance the STEM labor force, research and innovation, enhance the economy, and reduce the risk of women's social exclusion to the benefit of society.

The European Union's Strategy and Initiatives for Gender Equality in STEM

The European Commission objectives. Over the years, the EC has developed a regulatory framework on gender equality targeting the labor market and research with three main objectives: gender equality in careers, gender balance in decision-making bodies, and integration of the gender dimension in R&I. Concomitantly, gender equality and mainstreaming are among the Priorities of the European Research Area (ERA),^f while Article 16 of the Framework Regulation mandates the effective promotion of gender equality and the inclusion of the gender dimension in the R&I content. Thus, gender equality in Horizon 2020^g (H2020) is both a quantitative

c https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality_en

d <https://bit.ly/2Bg77On>

e <https://eige.europa.eu/rdc/thesaurus/terms/1185>

f http://ec.europa.eu/research/era/index_en.htm

g <https://ec.europa.eu/programmes/horizon2020/en/>

a http://ec.europa.eu/research/science-society/document_library/pdf_06/structural-changes-final-report_en.pdf

b <https://eige.europa.eu/rdc/thesaurus/terms/1168>

Within the EU, gender equality in all aspects of social, political, and cultural life, including education and R&I, is approached as a matter of social justice and fairness.

(for example, gender balance in research teams, evaluation panels, advisory boards, expert groups, and so forth) and a qualitative mandate, that is, inclusion of the gender dimension in research.

The EC Advisory Group on Gender in its December 2016 position paper differentiates inclusion of the gender dimension in research from gender balance, which is constituted as "... a dynamic concept that entails researchers taking into account sex and gender in the whole research process, when developing concepts and theories, formulating research questions, collecting and analyzing data using the analytical tools that are specific to each scientific area."^h The same document provides concrete advice on implementing the gender dimension in research for each H2020 Work Programme including Leadership in Enabling and Industrial Technologies. Conceptually, the EC research framework moves beyond the numerical, sometimes token, inclusion of women in research and ensures that gender and the way it impacts research and its outcomes are meaningfully taken into consideration. On the implementation level, in the EU, Gender Equality bodies, agencies, and associations, both EC and non-EC related, have been founded, and initiatives such as prizes and awards and specifically dedicated funding schemes have been established.

Indicative gender equality-related bodies, agencies, and associations in Europe. The European Institute for Gender Equality (EIGE) is an autonomous body of the EU with the goal to strengthen and promote GE, including gender mainstreaming in all EU and the resulting national policies. EIGE has developed the Gender Equality in Academia and Research (GEAR) Tool that provides a step-by-step guide to preparing GE plans for academic and research organizations. The Helsinki Group on Gender Equality in Research and Innovation, a Standing Working Group of the ERA Committee, brings together representatives from Member States and Associated Countries to advise the European Commission on policies and

initiatives on GE in R&I. Since 2005, She Figures and its Statistical Correspondents have published tri-annually pan-European comparable statistics on the current state of GE in R&I, thus serving the crucial goal of monitoring the progress toward GE and the impact of related policies and initiatives.

Focusing specifically on STEM and ICT, the European Centre for Women and Technology (ECWT) is a multiple stakeholder partnership consisting of more than 130 organizations and a significant number of individuals from governments, business, academia, and non-profit sectors with high-level expertise in women in technology development. It aims at increasing the number of girls and women in STEM and integrating a critical mass of women in the design, research, innovation, production, and use of ICT in Europe. Additionally, the European Network for Women in Digital aims at enhancing women's participation in digital studies and occupations across the EU.

An indicative list of non-EC related bodies, agencies, and associations promoting GE issues in the ICT field in Europe is presented in a report by Informatics Europe.⁹ Among them, the Athena-SWAN Charter promotes practices to eliminate gender bias and foster an inclusive culture that values female staff, partially through the establishment of prizes and awards. It has been identified as a most effective approach since approximately 82% of U.K. research institutions have adapted their strategies to its Charter scheme.¹⁰

Prizes and awards. The EU Prize for Women Innovators (<https://bit.ly/2tfcrxq>) is awarded every year to European women who founded a successful company and brought an innovation to market. The EC Call for Tech StartUps recognizes women who co-own a tech startup. Departments or faculties of EU universities or research institutes and labs that demonstrate a positive impact on women may be candidates for the MINERVA Informatics Europe Equality Award. European women in STEM may also apply for awards of international scope that recognize STEM-related achievements or for (European or international) STEM-related awards that target both genders.

^h <https://bit.ly/2Tu8dgA>

EC funding opportunities and funded projects. GE issues and the gender dimension in research constitute a crosscutting priority in the entire H2020 Work Programme. Nevertheless, a dedicated funding scheme is included in the H2020 Science with and for Society (Swafs) program.ⁱ Swafs projects contribute to the promotion of Gender Equality Plans (GEPs): A set of actions aimed at conducting impact assessment/audits of procedures and practices to identify gender bias, identifying and implementing innovative strategies to correct bias, and setting targets and monitoring process via indicators.^j Implementation of the respective GEAR Tool has resulted in examples of best practices on how to attract women into academic leadership positions ensuring, for instance, a gender-balanced representation in the highest decision-making bodies of universities.^k Nonetheless, Swafs represents only 1.5% of the total budget for all activities under the Societal Challenges section.⁸

With respect to GE in STEM, a significant proportion of EC-funded projects are aimed at structural changes in science, technology, and innovation research organizations and at the inclusion of the gender dimension in research and education. Such projects include (indicatively) GENERA (<https://genera-project.com/index.php>); GEECO (<http://www.gecco-project.eu/home>), which will set up GEPs for universities and funding organizations in the STEM area; and EFFORTI (<https://www.efforti.eu/>), aiming at analyzing the influence of measures to promote GE on R&I outputs and on establishing more responsible and responsive research, technology, development, and innovation systems.

A flagship project of particular interest to STEM is Gendered Innovations (<http://genderedinnovations.stanford.edu>), which specifically addresses the gender dimension of R&I. The project has developed practical

methods of gender analysis tailored to the needs of scientists and engineers. More importantly, the project provides peer-reviewed analyses of case studies that evidence the need for considering gender in all stages of research design and implementation in order to produce better science and innovation outputs.

Expected Impact

Measures and initiatives to promote GE in STEM fields have shown positive effects. In Germany, for instance, education initiatives contributed to an increase in the number of women graduating in STEM-related fields.¹ Similarly, women's share among appointed STEM professors has increased by 4.1%.^{2,11} Dedicated funding further contributes to an increasing interest in GEP implementation among EU research institutions and organizations (113 organizations through 17 projects up to 2017).

Closing the gender gap in STEM is further expected to increase the scientific quality and societal relevance of produced knowledge, technologies, and innovations; contribute to the production of goods and services better suited to potential markets;¹ and further aid EU economic growth. For example, it is estimated that by 2030 the increase of women's participation in STEM-related fields will increase the EU GDP per capita by 0.7–0.9.⁶ In monetary terms, this will lead to 610–820 billion euros improvement in GDP. Furthermore, if effectively implemented, relevant EC measures are expected to increase women's employment, productivity, and wages⁴ and thus contribute to long-term competitiveness of the EU economy and improved balance of trade.⁶

Conclusion

Narrowing the gender gap in STEM fields has the potential to increase European labor supply and market activity, make women (and men) better equipped to secure steady and well-paid jobs, and in turn reduce the risk of women's social exclusion, improving both science and society as a whole.

References

- Anger, C. et al. MINT—Frühjahrssreport (STEM-Spring report). Institut der deutschen Wirtschaft Köln (Ed.). Institut der deutschen Wirtschaft Köln, Germany, 2012.
- Bund-Länder-Kommission für Bildungsplanung und Forschungsförderung. Frauen in Führungspositionen an Hochschulen und außerhochschulischen Forschungseinrichtungen: Zehnte Fortschreibung des Datenmaterials (Women in leading positions at universities and non-university research organizations: 11th update to the data). Materialien sur Bildungsplanung und zur Forschungsförderung, 139, 2005.
- Burchell, B. et al. A New Method to Understand Occupational Gender Segregation in European Labour Markets. Publications Office of the European Union; https://ec.europa.eu/info/sites/info/files/150119_segregation_report_web_en.pdf
- Caprile, M. Encouraging STEM studies—Labour Market Situation and Comparison of Practices Targeted at Young People in Different Member States, European Parliament, DG for Internal Policies. Policy Department A, 2015.
- Dobson, I. STEM; Country Comparisons—Europe. A critical evaluation of existing solutions to the STEM skills shortage in comparable countries. Australian Council of Learned Academies, 2013.
- EIGE. Economic benefits of gender equality in the EU. How gender equality in STEM education leads to economic growth (2018); <https://eige.europa.eu/rdc/eige-publications/economic-benefits-gender-equality-eu-how-gender-equality-stem-education-leads-economic-growth>
- European Centre for Development of Vocational Training. Skill shortage and surplus occupations in Europe. Briefing Note (Oct. 2016); http://www.cedefop.europa.eu/files/9115_en.pdf
- European Commission. Interim Evaluation: Gender equality as a crosscutting issue in Horizon 2020, DG for Research and Innovation, Directorate B—Open Science and Open Innovation, 2017; https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/interim_evaluation_gender_long_final.pdf
- Informatics Europe. Informatics Education in Europe: Institutions, degrees, students, positions, salaries, Key data 2012–2017 (Oct. 2018).
- Kalpazidou, E. Schmidt, et al. A Conceptual evaluation framework for promoting gender equality in research and innovation. Toolbox I – A synthesis report. Deliverable 3.3 EFFORTI Project; <https://bit.ly/2t74pWM>
- Kathinka, L. et al. Gender and STEM in Germany: Policies Enhancing Women's Participation in Academia, International Journal of Gender Science and Technology, 2012; <http://genderandset.open.ac.uk/index.php/genderandset/article/view/304>
- UNESCO Institute for Statistics Women in Science Fact Sheet No. 51, June, FS/2018/SCI/51; <http://uis.unesco.org/sites/default/files/documents/fs51-women-in-science-2018-en.pdf>
- Wragg, J. A new style of learning is essential to plugging the STEM skills gap (2016); <http://www.itportals.com/2016/01/03/new-style-learning-essential-plugging-stem-skills-gap/#ixzz3xp01vBU4>.
- Yang, X., Gao, C. Missing women in STEM in China: A micro-level explanation for achievement motivation and gender socialization. Oral presentation XIX ISA World Congress of Sociology. Toronto Canada, July 15–21, 2018; <https://isaconf.confex.com/isaconf/wc2018/webprogram/Paper97933.html>

Panagiota Fatourou is an associate professor at the University of Crete, Department of Computer Science and a collaborating faculty member of the Foundation for Research and Technology-Hellas (FORTH), Institute of Computer Science (ICS), Greece.

Yota Papageorgiou is a professor at the University of Crete Department of Sociology, Greece.

Vasiliki Petousi is an assistant professor at the University of Crete Department of Sociology, Greece.

ⁱ A complete list of the projects funded by the FP7 and H2020 programs is provided at <https://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=gender>

^j <https://eige.europa.eu/gender-mainstreaming/toolkits/gear/what-gender-equality-plan-gep>

^k Selected abstracts of best practices can be found <https://bit.ly/2Gq9q4E>.

¹ <https://eige.europa.eu/gender-mainstreaming/policy-areas/research>

DOI:10.1145/3310330

BY MICHAEL E. CASPERSEN, JUDITH GAL-EZER,
ANDREW MCGETTRICK, AND ENRICO NARDELLI

Informatics as a Fundamental Discipline for the 21st Century

INFORMATICS FOR ALL is a coalition whose aim is to establish informatics as a fundamental discipline to be taken by all students in school. Informatics should be seen as important as mathematics, the sciences, and the various languages. It should be recognized by all as a truly foundational discipline that plays a significant role in education for the 21st century.

The European scene. In Europe, education is a matter left to the individual states. However, education, competencies, and preparedness of the workforce are all important matters for the European Union (EU).

Importantly, there is a recognition that the education systems of Europe do not collectively prepare students sufficiently well for the challenges

of the digital economy. These systems need to be fundamentally transformed and modernized. In January 2018, a Digital Education Action Plan,¹ which set out a number of priorities, was published by the EU. The most relevant priority for our initiative is “Developing relevant digital competences and skills for the digital transformation,” and the Plan suggests one way to implement this is to “Bring coding classes to all schools in Europe.” This is important, but more is needed, as we will explain in this article.

ACM Europe and Informatics Europe. ACM Europe (europe.acm.org) was established in 2008, and Informatics Europe (www.informatics-europe.org) in 2006. From the early days, the two organizations have collaborated on educational matters; through this liaison, they are seen to project to the wider community a single message about aspects of informatics^a education. In 2013, the two groups set up and funded a Committee on European Computing Education (CECE) to undertake a study that would capture the state of informatics education across the administrative units of Europe (generally, these units are the countries, but within Germany, for instance, there are 14 different administrative units with autonomy regarding education).

The CECE study paralleled the highly influential U.S. study *Running on Empty*¹¹ that had drawn attention to the state of computer science education in the U.S. The CECE study gathered data from 55 administrative units (countries, nations, and regions) of Europe (plus Israel) with autonomous educational systems through the use of questionnaires and a wide network of reliable contacts and official sources.

The report on that work was published in 2017.³ The three themes of informatics, digital literacy, and teacher training provided the framework for the study. Informatics was

^a In most of Europe, informatics is synonymous with computing or computer science.

INFORMATICS



defined as knowledge and competencies about computational structures, processes, artifacts, and systems. Digital literacy was seen as basic user skills, such as conversancy with standard tools like word processors, Web browsers, spreadsheets, and so on.

While the report confirmed that, across Europe, there was a growing realization of the importance of sound school education in informatics, it also showed a largely variable level of effort and achievement across administrative units. For instance, the report found that informatics was available to all pupils in only 22 out of 50 units, while in a further 10 units it was available to just some students, and in several noticeable cases, no informatics

teaching was available at all. When students could elect to take an informatics course, there was evidence of poor uptake, often as low as 10%.

The authors of the CECE report included a number of recommendations that would serve to improve the situation. Those recommendations addressed each of the three areas (informatics, digital literacy, and teacher training), and these form the basis of the Informatics for All initiative.

Informatics for All

The task of moving forward with the CECE recommendations was seen as different in character from the survey work. Importantly, the Informatics for All Coalition was formed

by the joint efforts of ACM Europe, Informatics Europe, and the Council of European Professional Informatics Societies.⁴ These organizations all share a common concern about the state of informatics education throughout Europe, and are committed to stimulating activity that will lead to significant improvement.

In moving forward, the new organization took the opportunity to present a perspective on informatics education that would reflect the advances that have occurred since 2014, when the CECE work properly began.

Building on the CECE recommendations, the report *Informatics for All: The Strategy*¹² was produced. The emphasis of the report is on informatics education, with informatics seen as the science underpinning the development of the digital world—a distinctive discipline with its own scientific methods, its own ways of thinking, and its own technological developments.

By emphasizing the constructive and creative elements of the discipline, the role of informatics in innovation and discovery and its role in shaping the digital world, the discipline is seen as an essential element of education for the 21st century. Its role in competitiveness and in the economic prosperity of Europe (and beyond) further adds to its vital nature.

The report, which contains eight recommendations (see the accompanying figure), was formally launched in Brussels in March 2018. The launch was attended by representatives of the European Commission as well as representatives of industry and academia; it received uniform, enthusiastic support.

Two-tier strategy. In many ways, the Informatics for All initiative mirrors the CS for All initiative launched in the U.S. in early 2016 (see sidebar). A crucial element of the European approach, which distinguishes it from the CS for All initiative, is the two-tier strategy at all educational levels: Informatics as an area of specialization, that is, a fundamental and independent subject in school; and the integration of informatics with other school subjects, as well as with study programs in higher education. Perhaps overly simplified, the two tiers

The eight recommendations from *Informatics for All: The Strategy*.¹²

Curriculum Considerations

- ▶ All students must have access to ongoing education in informatics in the school system. Informatics teaching should start in primary school.
- ▶ Informatics curricula should reflect the scientific and constructive nature of the discipline, and be seen as fundamental to 21st century education by all stakeholders (including educators, pupils, and their parents).
- ▶ Informatics courses must be compulsory and recognized by each country's educational system as being at least on a par with courses in STEM (Science, Technology, Engineering, and Mathematics) disciplines. In particular, they must attract equivalent credit, for example, for the purposes of university entrance.

Preparing Teachers

- ▶ All teachers at all levels should be digitally literate. In particular, trainee teachers should be proficient (via properly assessed courses) in digital literacy and those aspects of informatics that support learning.
- ▶ Informatics teachers should have appropriate formal informatics education, teacher training, and certification.
- ▶ Higher education institutions, departments of education, as well as departments of informatics should provide pre-service and in-service programs, encouraging students to enter a teaching career related to informatics.
- ▶ Ministries should be encouraged to establish national or regional centers facilitating the development of communities of informatics teachers who share their experiences, keep abreast of scientific advances, and undertake ongoing professional development.

Teaching the Teachers

- ▶ Intensive research on three different facets, curriculum, teaching methods and tools, and teaching the teachers is needed to successfully introduce informatics into the school system.

The U.S. Initiative CS for All

The CS for All initiative, launched by President Barack Obama on January 30, 2016,⁶ was highly imaginative and a catalyst for a burst of initiatives in computer science (CS) education in the U.S. It fired the imagination and provided a focus for great activity centered on the promotion of CS at all stages of education. The financial commitments were impressive, even eye-watering!

The initiative could be seen as the culmination of earlier work on CS education supported by the National Science Foundation (NSF), the CS Principles course launched by the College Board,⁷ the extensive work of the Computer Science Teachers Association (CSTA),⁸ and efforts by ACM, by code.org, and by many others. Within ACM, the efforts included work on policy matters by the Education Policy Committee, harnessing the invaluable support of major industrial players through Computing in the Core (which has now merged with the code.org Advocacy Coalition), lobbying on Capitol Hill, as well as actions from groups with members in the ACM Education Council.

may be characterized as Learn to Compute (specialization) and Compute to Learn (integration).

All students—regardless of their special interest, area of expertise, and future profession—need to be educated in informatics and apply those knowledge and skills as an integrated competence in all subjects and professions. In so doing, they must ensure that technological development is directed towards the achievement of a better, safer, fairer, and more just society.

The second tier of the strategy, integrating informatics with other disciplines, has huge educational potential.

Through digital models, subjects can be taught in novel and more engaging ways, and data-driven approaches will open doors to new dimensions of understanding and new ways of learning subjects. Similarly, through programming of, say, simulations and games, knowledge and insight in a subject can be expressed in more individual, novel, useful, and creative ways (instead of the traditional reproduction of knowledge in written or oral form).

By integrating informatics in other disciplines, students are provided the advantage of having an additional novel, specific way of thinking to describe and explain phenomena (often referred to as “computational thinking”), complementing that of other scientific disciplines and contributing to their better, more thorough understanding. This is pursued even in STEM, for example, Weintrop,¹⁶ and K-12 SF.¹⁴

Implementation

The challenge now for the Informatics for All Committee is to bring about change leading to the realization of the strategy. It is highly unlikely that the recommendations will simply be mandated; rather, a carefully considered approach that leads toward the acceptance of the recommendations seems far more realistic.

The implementation problem has to be addressed within each country where responsibility for education resides. Within each country, groups consisting of administrators, academics, teachers, industrialists,

employers, and others, could come together and (with sensitivity and caution) create pressure and initiatives that would lead to change as required.

There are different areas of responsibility and different possible activities within these areas:

Education authorities. Administrators have responsibility for the proper recognition of disciplines, and related matters. Accordingly, they have a role in the implementation of certain aspects of the recommendations:

- recognition of informatics as a science,
- the education of teachers of informatics, and
- the education of all teachers.

Curriculum development (includes pupils and parents). Within the CECE report, there are comments that suggest current informatics curricula are not uniformly popular with pupils and their parents. For informatics to be a discipline taken by all, there is a need to review and revise (and in some cases, design) curricula to ensure the discipline is considered an essential one for the 21st century by all stakeholders, including students and their parents.

The motivation of students must be heightened dramatically; all students, including the best students, must see informatics as highly relevant. The curriculum is predominantly technical in nature, and has to capture the essence of the discipline, emphasizing the relevance to people and society in general and to the young in particular, thus including fundamental issues with the practical and more theoretical aspects being carefully interwoven. To motivate students, attention can be drawn to creativity, innovation, and applications, and the massive impact these have on society, particularly highlighting the use of big深深 data, Internet of Things, and developments in machine learning and their impact on the ‘future of work.’

Role of higher education. Within higher education institutions (HEI), expertise should be mobilized to support the development of competence and capacity. There are four main aspects:

The emphasis of the report is on informatics education, with informatics seen as the science underpinning the development of the digital world—a distinctive discipline with its own scientific methods, its own ways of thinking, and its own technological development.

The challenge now for the Informatics for All Committee is to bring about change leading to the realization of the strategy.

► Staff in HEIs can provide expertise and advice to guide developments, such as in advising on relevant curricular standards (detailing what can and should be taught at different stages of school education).

► HEIs often are engaged in the education and training of teachers, and in their professional development. Their forward thinking can serve to continually improve teaching expertise; this can happen at the stage of initial teacher education, and through continuing professional development.

► Within HEIs, research typically features strongly, and the Informatics for All strategy stresses the importance of research. In this context, HEIs might partner with school-teachers, educationists, and/or others to drive forward relevant research agendas.

► HEIs are a powerful force, perhaps the single most powerful force, in terms of influencing the delivery of the curriculum in schools. What are the requirements for entry to specific programs of study?

Role of professional bodies and the EU. Professional bodies, such as the Council of European Professional Informatics Societies (CEPIS), will typically have strong links with industry and will be in a position to effectively harness industrial perspectives; they typically will be recognized as such by government. It is expected that they will voice strong views about the need for informatics education, especially in relation to the economic development of the country and workforce planning. Where their feelings are particularly strong, they can bring pressure to bear at the governmental level to provide national resources for teacher training, research, and more. They also may be able to provide resources to aid education. It would also be desirable to complement the EU's Digital Action Plan with national initiatives, possibly supported by EC resources.

Curriculum issues. Adopting a completely new subject to a national school curriculum is challenging for many reasons. A general but very concrete challenge is how to provide the necessary space in the curricula. We do not believe that the school system

of 2019 has reached a fixed point with respect to mandatory subjects; thus, in whichever way possible, space must be found. We feel that each country will have to find her own solution, matching her constraints and situation.

As mentioned, a curriculum should include the foundations of the discipline, including theoretical and implementation aspects. Clearly it should not be just a downgrade of university curricula, but a curriculum should be specially designed for each school level (elementary, middle, and high school). New curricula research should be conducted to examine and find the appropriate methodology and pedagogy to design such curricula for the different levels of school.

Human and Societal Perspectives

Informatics dramatically differs from other sciences in terms of the way it empowers. With informatics, there is powerful support for automating cognitive tasks¹⁵ and this has implications for all domains and professions. Moreover, the related concept of computational thinking is recognized as having relevance more widely.^{2,17,18}

In Iversen,¹³ the concept of computational empowerment is seen as an important development of computational thinking that places an emphasis on the abilities needed to effectively deploy informatics. Based on a critical analysis of current tech-

Informatics for All The strategy

ACM Europe & Informatics Europe
February 2018



nologies, it pulls together technological insights and innovation and links these to the role that informatics plays in the development of personal life and of society. In this way, future generations will have the knowledge and skills from informatics to become competent, constructive, and critical co-creators of the digital future.

In a context where informatics education begins early in primary school and is carried forward through to the later years of high school, there is the opportunity to develop thoughts about the possible wider relevance of ideas from informatics, and to develop them in a meaningful way that places emphasis on the human (and societal) benefits and implications. Just making suggestions about possible improvements opens the door to considerations about (software-inspired) innovation, and more.

It is important that students acknowledge software as creator and bearer of values and culture—that these aspects are explicitly or implicitly embedded in the software. Software is formed through design processes that include critical decision-making; students must learn to creatively develop software, and learn to analyze and understand the impacts of software and digital artifacts in general. For example, these visions for a strong human and societal perspective are thoroughly embedded in the newly designed Danish curriculum for informatics in school.⁹

Augmented intelligence. The concept of augmented intelligence relates to the effective use of informatics in augmenting human intelligence. The discussion above regarding STEM was one instance of that though even there, there is scope for extending that further.

Developments in language translation, voice recognition and simulation, and related advances fueled by developments in machine learning suggest great scope for reshaping the curriculum in many disciplines.

The role of ethics. Those who develop software ought to do so in a responsible manner, ensuring that ‘bad things’ do not happen. The related issues tend to be captured in a code of conduct that provides guidance on

good practice. In the past, such codes have tended to stress matters such as ‘do not cause harm’. While this remains important, a more enlightened approach places an emphasis on contributing positively to the benefit of a fair, just, and secure society through the use of computers and computing. The recent ACM Code of Ethics and Professional Conduct⁵ takes such a view.

The role of teachers. Teachers are the key to the success of the implementation of any study program and the introduction of any new curriculum or technologies. A good supply of well-educated and enthusiastic teachers is crucial to support every discipline in schools at all levels, but the lack of suitable teachers at all levels also forms a bottleneck for the successful implementation of Informatics for All. Thus, efforts should be devoted to recruitment, and to establishing a supportive informatics teacher community.

Concluding Remarks

The primary focus of Informatics for All is to stimulate the recognition that informatics is a vital, important discipline, both as a subject of study on its own, and also integrated with other disciplines with many of the ideas having relevance more broadly.

The two-tier approach facilitates the integration of informatics into the teaching of other disciplines, reshaping the curriculum for all disciplines and generally providing a basis for making education systems truly relevant for the 21st century. It also opens up many avenues for research; for instance, about how to teach disciplines effectively in a world of constant change.

Given that digital technology is taking an increasingly relevant and pervasive role, providing to all citizens an appropriate level of informatics education is necessary to ensure balanced development of the digital society.

Informatics for All is a catalyst for important thoughts for reforming the wider educational systems to the benefit of students and employers, and ultimately the economy of Europe.

Acknowledgments. The authors wish

to acknowledge contributions and support from various quarters: Wendy Hall, Bob McLaughlin, Bobby Schnabel and the board members of ACM Europe, Informatics Europe, and the original CECE team. □

References

1. Digital Action Plan, The European Commission, Brussels, Jan 17, 2018; <https://ec.europa.eu/education/sites/education/files/digital-education-action-plan.pdf>
2. Aho, A.V. Computation and computational thinking. *Ubiquity* (Jan. 2011), article 1. ACM Press. DOI: <https://doi.org/10.1145/1922681.1922682>
3. Informatics Education in Europe: Are we all in the same boat? The Committee for European Computing Education. ACM/Informatics Europe, 2017; <https://doi.org/10.1145/3106077>, 2017.
4. cepis.org
5. ACM Code of Ethics and Professional Conduct, July 2018, <https://www.acm.org/code-of-ethics>
6. Computer Science For All, The White House, <https://obamawhitehouse.archives.gov/blog/2016/01/30/computer-science-all>.
7. <https://apcentral.collegeboard.org/courses/ap-computer-science-principles>
8. Computer Science Teachers Association, <https://www.csteachers.org>
9. Curriculum for ‘teknologiforståelse’: Danish Ministry of Education; <https://bit.ly/2LswHUF>
10. e-Competence Framework (e-CF)—A Common European Framework for ICT Professionals in All Industry Sectors. Published as EN 16234-1:2016, 2016.
11. Wilson, C., Sudol, L.A., Stephenson, C. and Shetlik, M. *Running on Empty: The Failure to Teach K-12 Computer Science in the Digital Age*. Published by ACM and CSTA, 2010.
12. Caspersen, M.E., Gal-Ezer, J., McGettrick, A. and Nardelli, E. *Informatics for All: The Strategy*. Produced on behalf of ACM Europe and Informatics Europe. ACM, New York, Feb. 2018.
13. Iversen, O.S., Smith, R.C. and Dindler, C. From computational thinking to computational empowerment: A 21st century PD agenda. In *Proceedings of PDC 18*, Hasselt and Genk, Belgium.
14. National Research Council. A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academic Press.
15. Nardelli, E. The maintenance is the implementation OR Why people misunderstand IT systems. In *Proceedings of the 6th European Computer Science Summit* (Prague, Czech Republic, Oct. 2010); <https://bit.ly/2M02KVB>
16. Weintrop, D. et al. Defining computational thinking for mathematics and science classrooms. *J. Science Education and Technology* 25, 1 (Oct. 8, 2015), 127–147.
17. Wing, J. Computational thinking. *Commun. ACM* 49, 3 (Mar. 2006), 33–35.
18. Wing, J. Computational Thinking—What and Why? The Link. *News from the School of Computer Science at Carnegie Mellon University*, 2011.

Michael E. Caspersen is managing director of It-vest – networking universities, and honorary professor at Aarhus University, Aarhus, Denmark.

Judith Gal-Ezer is professor emerita of The Open University of Israel, Ra'anana, Israel and vice chair of ACM Europe.

Andrew McGettrick is professor emeritus of Strathclyde University, Glasgow, Scotland.

Enrico Nardelli is a professor of informatics at Tor Vergata University of Rome, Italy, and president of Informatics Europe.

BY PAOLA INVERARDI

The European Perspective on Responsible Computing

WE LIVE IN the digital world, where every day we interact with digital systems either through a mobile device or from inside a car. These systems are increasingly autonomous in making decisions over and above their users or on behalf of them. As a consequence, ethical issues—privacy ones included (for example, unauthorized disclosure and mining of personal data, access to restricted resources)—are emerging as matters of utmost concern since they affect the moral rights of each human being and have an impact on the social, economic, and political spheres.

Europe is at the forefront of the regulation and reflections on these issues through its institutional bodies. Privacy with respect to the processing of personal data is recognized as part of the fundamental rights and freedoms of individuals. Regulation (EC)

45/2001 establishes the rules for data protection in the EU institutions and the creation of the *European Data Protection Supervisor* (EDPS) as independent supervisory authority to monitor and ensure people's right to privacy when EU institutions and bodies process their personal data. The *European Group on Ethics in Science and New Technologies* (EGE) is an independent advisory body of the President of the European Commission that advises on all aspects of Commission policies and legislation where ethical, societal, and fundamental rights dimensions intersect with the development of science and new technologies. In 2015, the EDPS appointed the Ethics Advisory Group (EAG) "to explore the relationships between human rights, technology, markets, and business models in the 21st century."

Autonomous systems. We broadly define autonomous systems as systems that have the ability of substituting humans in supplying (contextual) information that the system may use to make decisions while continuously running. Depending on the nature, property, and use of this information, an autonomous system may impact moral rights of the users, be they single citizens, groups, or the society as a whole. The widespread use of AI techniques in the implementation of these systems has exacerbated the problem contributing to the creation of systems and technologies whose behavior is intrinsically opaque.^{1,2,14} In this article, we will stick to the notion of autonomous technology rather than with AI technology. Indeed, we are concerned with the autonomous decision-making capabilities of systems even if those capabilities are a consequence of the availability of more and more complex AI enabling technologies.

The harm of digital society. The last years have witnessed an increasing rate of concerns on the impact of autonomous technologies on our societies. Economy, politics, and human being natural rights are endan-



The GDPR aims to give individuals control over their personal data and to provide a unifying regulation within the EU for international business.

gered by the uncontrolled use of autonomous technology. Institutional as well as social and scientific entities and boards contribute to constantly feeding the debate by advocating and proposing codes of ethics for developers and regulations from governmental bodies.^{1-3,12-15,21,22} Admittedly, this debate is mostly concentrated in western countries although with different regulatory outcomes. Indeed, ethical principles, notably privacy, may vary from country to country due to their specific culture and history^{16,17} and to the impact the development of autonomic technologies can have on the economy of the country. However, at least in western countries there is growing consensus that it is time to take actions to address the harms of autonomous technologies¹⁵ and that those actions need eventually to have a regulatory nature and be part of public policy.^{18,19} To this respect, Europe is certainly far ahead both in thinking and regulation.

The General Data Protection Regulation (GDPR), which is the most advanced in the world regulation on personal data protection, is Europe's most relevant achievement so far. By comparison, the state of California recently passed a digital privacy law that will go into effect in January 2020. Although more limited in scope than GDPR, the law is considered one of the most comprehensive in the U.S.²⁰ In a recent paper, "Constitutional Democracy and Technology in the Age of Artificial Intelligence,"¹⁹ Paul Nemitz, Principal Advisor of the European Commission, claimed that "The EU GDPR is the first piece of legislation for AI." He provides a comprehensive account of the debate and of the process that accompanied the formulation and adoption of GDPR. Nemitz points out that as happened with GDPR concerning personal data protection, AI and autonomous technologies need to be regulated by laws as far as individual fundamental rights and democracy of society are concerned.

This would lead to accept AI-based autonomous technologies only "if by design, the principles of democracy, rule of law, and compliance with fundamental rights are incorporated in AI, thus from the outset of program

development," Nemitz writes.

The quest for an ethical approach. For years, Europe has called for a more comprehensive approach that encompasses privacy and addresses ethical issues in the scope of the digital society. The EDPS in its strategy for 2015–2019 sets out the goal to develop an ethical dimension to data protection.⁴ In order to reach the goal, it has established the EAG with the mandate to steer a reflection on the ethical implications that the digital world emerging from the present technological trends puts forward. EDPS 4/2015 Opinion "Toward a new digital ethics,"³ identifies the fundamental right to privacy and the protection of personal data as core elements of the new digital ethics necessary to preserve human dignity as stated in Article 1 of the EU Charter of Fundamental Rights. The Opinion also calls for a big data protection ecosystem that shall involve developers, businesses, regulators, and individuals in order to provide 'future-oriented regulation,' 'accountable controllers,' 'privacy-conscious engineering,' and 'empowered individuals.'

In its 2018 report,⁶ the EAG has provided a broader set of reflections on the notion of digital ethics that address the "fundamental questions about what it means to make claims about ethics and human conduct in the digital age, when the baseline conditions of humanness are under the pressure of interconnectivity, algorithmic decision-making, machine-learning, digital surveillance, and the enormous collection of personal data." In March 2018, the EGE released a statement on "artificial intelligence, robotics, and 'autonomous' systems" in which it urges an overall rethinking of the values around which the digital society is to be structured.⁵ Computer scientists, besides other societal actors, are called to join this effort by contributing theories, methods, and tools to build trustable and societal-friendly systems. "Advances in AI, robotics and so-called 'autonomous' technologies have ushered in a range of increasingly urgent and complex moral questions," the EGE states. "Current efforts to find answers to the ethical,

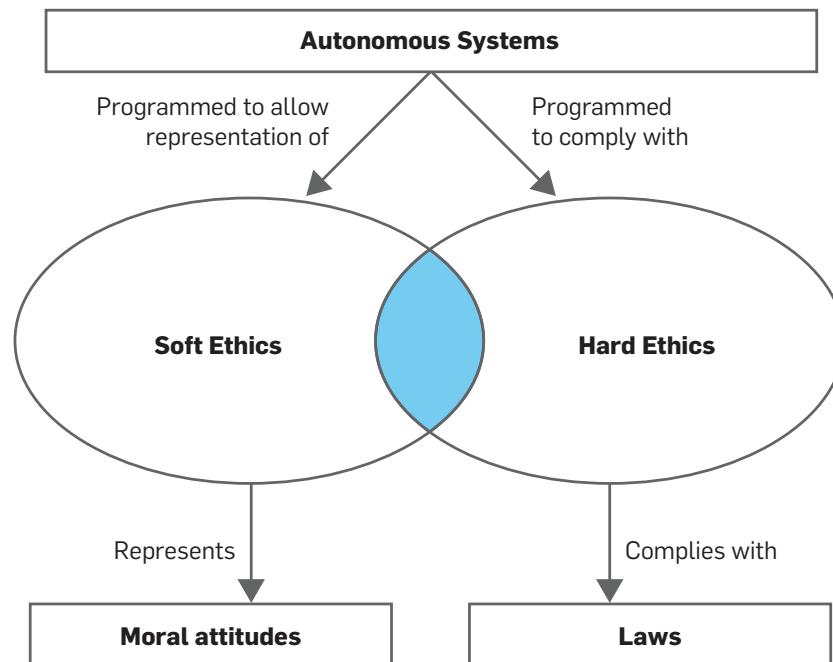
societal, and legal challenges that they pose and to orient them for the common good represent a patchwork of disparate initiatives. This underlines the need for a collective, wide-ranging, and inclusive process of reflection and dialogue, a dialogue that focuses on the values around which we want to organize society and on the role that technologies should play in it."

In its statement, the EGE goes further and proposes "a set of basic principles and democratic prerequisites, based on the fundamental values laid down in the EU Treaties and in the EU Charter of Fundamental Rights." The first one restates human dignity in the context of the digital society: "(a) Human dignity, The principle of human dignity, understood as the recognition of the inherent human state of being worthy of respect, must not be violated by 'autonomous' technologies."

Supporting ethical concerns in autonomous systems. Europe is thus calling for a digital society in which the human being with fundamental rights remains at its center. Therefore, there is the need to rethink the role of the various actors in the digital world by empowering the users of the digital technology both when they operate as citizens and as individuals. However, what does it mean to empower the citizens and the individuals?

Human at the center. The stated principle of human dignity indicates that individuals need to be able to exercise some degree of control on their information and on the decisions that autonomous systems make on their behalf. This raises an issue of what is the scope of system autonomy. Indeed, the principle asks for autonomous systems that, in their behavior, pay respect to human's decisions and beliefs. This means that system's autonomy is a direct consequence of the amount and kind of respect of the individuals they interact with. The more individuals the system interacts with the less autonomy may be given to potential conflicts of respect. This is clearly understood in the scope of privacy where different individuals may have different privacy concerns about their personal data

A conceptualization of the relationship between digital ethics and autonomous systems.



both in general and also depending on given contexts. Reflections on digital ethics can help us in shaping the scope of system autonomy.

Digital ethics. Luciano Floridi, a professor of philosophy and the ethics of information at Oxford and director of the Digital Ethics Lab of the Oxford Internet Institute, defines digital ethics⁷ as the branch of ethics that aims at formulating and supporting morally good solutions through the study of moral problems relating to personal data, AI algorithms, and corresponding practices and infrastructures. Simplifying, he further identifies two separate components of digital ethics, hard and soft ethics. Hard ethics is defined and enforced by legislation. However, legislation is necessary but insufficient, since it does not cover everything, nor should it. In the space that is left open by regulation, the actors of the digital world, for example, companies, citizens, and individuals, should exploit digital ethics in order to forge and characterize their identity and role in the digital world. This is the domain of soft ethics, which deals with what ought and ought not to be done over and above the existing regulation,

without trying to bypass or change the hard ethics.

From the user perspective, soft ethics is where individual ethical values can be expressed; hard ethics characterizes the values, defined by the legislation, a digital system producer shall comply with. Soft ethics is therefore the context in which a user's control of autonomous technology shall and can be exercised.

A patchwork of approaches. Besides reflections and statements on ethics, Europe has put in place a number of initiatives that on the one side represent a patchwork according to the EGE, and on the other side they show that ethical concerns are at the core of the interest for the European society at a whole. A few examples follow:

From a regulatory standpoint, the GDPR was entered into application throughout the EU in May 2018. Article 1 states that: "Regulation lays down rules relating to the *protection of natural persons* with regard to the processing of personal data and rules relating to the free movement of personal data. This Regulation protects *fundamental rights and freedoms of natural persons* and in particular

Users need to be able to verify the system they use by possibly imposing on them their own ethical requirements.

their *right to the protection of personal data.*”

The GDPR aims to give individuals control over their personal data and to provide a unifying regulation within the EU for international business. It states data protection rules for all companies operating in the EU, whether they are established in the EU or just operating inside the EU. This regulation forces controllers of personal data to shape their organization and their processing systems in order to implement the data protection principles. As already mentioned, GDPR is the most advanced regulation about personal data operating in the world.

Through its organizations, the scientific community has contributed (at a policy level) to identify problems and establish criteria to develop algorithms and systems that embed machine-learning-fueled autonomous capabilities. In March 2018, the ACM Europe Council, the ACM Europe Policy Committee (EUACM), and Informatics Europe presented a white paper on “When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making.”² The report critically analyzes the implications of the increasing adoption of machine learning automated decision making in modern autonomous systems. It concludes with a set of recommendations for policy-makers that concern the technical, ethical, legal, economic, societal, and educational dimensions of the digital society.

DECODE is a consortium of 14 European organizations (municipalities, companies, research institutions, foundations) led by the municipality of Barcelona. DECODE is developing a project, funded by the European Commission through its research programs Horizon 2020,⁸ whose aim is to empower citizens to control their personal information over the Internet. It provides a distributed platform and tools that use blockchain technology with attribute-based cryptography to give people control of how their data is accessed and used. DECODE experiments through pilots deployed in Amsterdam and Barcelona. The pilots focus on the Internet of Things,

collaborative economy and open democracy. The DECODE project was selected in response to a call that stated the following objective: “The goal is to provide SMEs, social enterprises, industries, researchers, communities and individuals with a new development platform, which is intrinsically protective of the digital sovereignty of European citizens.”

Beyond privacy, in particular regarding the potential conflict between user/social ethical principles and autonomous systems decisions, ethical issues insistently emerged in the autonomous cars domain. Indeed, there is no general consensus on which ethical principles (personal ethics settings versus mandatory ethics setting) need to be embedded, and how, in the control software of autonomous vehicles.^{9,10} In 2016, the German Federal Ministry of Transport and Digital Infrastructures appointed an ethical committee that produced a recommendation report resulting in 20 ethics rules for automated and connected vehicular traffic.¹¹ In particular, rules 4 and 6 mention the ethical principle of safeguarding the freedom of individuals to make responsible decisions and the need to balance that with the freedom and safety of others.

Challenges for computer scientists. Responsible computing as defined in the European perspective sets out a number of ambitious challenges for computer scientists.

When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making

Informatics Europe & EUACM
2018



ACM Europe
Policy Committee

Empowering the user requires a complete rethinking of the role of the user in the digital society. The user is no longer a passive consumer of digital technologies and a data producer for them. Her dignity as a human being implies ownership of personal data and freedom of making responsible decisions. Autonomous technologies shall be designed and developed to respect it. This lifts the user to become an independent actor in the digital society able to properly interact with the autonomous technologies she uses every day and equipped with the appropriate digital means.

The separation of digital ethics in hard and soft ethics suggests that hard ethics is what the autonomous system shall comply with while soft ethics is specific to each individual/user. To obey the principle of human dignity the system during its interactions with each individual shall not violate her soft ethics. The autonomous system architecture shall permit this interaction to happen by complying with the user's moral prerogatives and capabilities. Users need to be able to verify the system they use by possibly imposing on them their own ethical requirements. The separation of concerns implied by the above notion of digital ethics suggests an overall framework in which the autonomy of the system is delimited by hard ethics requirements, users are empowered with their own soft ethics, and the interactions between the system and each user are further constrained by their soft ethics requirements. Therefore, the capability of an autonomous system to make decisions does not only need to comply with legislation but also with a user's moral preferences. (See the intersection between soft and hard ethics in the accompanying figure.)

In such a framework, it should also be possible to deal with liability issues in a fine-grained way by distributing responsibility between the system and the user(s) according to hard and soft ethics. The envisioned framework requires several steps. On the ethics side, provided that autonomous systems will be developed by complying with hard ethics that is with the regulations, the crucial issue

to face is to respect each individual's soft ethics. If verifying the compliance of autonomous systems to hard ethics is already raising huge scientific interest and great worries (given the use of obscure AI techniques),^{1,2,14} defining the scope of soft ethics and characterizing individual ones is a daunting task. Indeed, neither a person nor a society applies moral categories separately. Rather, everyday morality is in constant flux among norms, utilitarian assessment of consequences, and evaluation of virtues. Nevertheless, a digital society that fully realizes the principle of human dignity shall allow each individual to express her soft ethics preferences. Further challenges concern means to consistently combine user soft ethics with system hard ethics and to manage interactions of the system with users endorsing different ethics preferences. Autonomous systems shall be realized by embedding hard ethics by design but remaining open to accommodate users' soft ethics. This could be achieved through system customization or by mediating interactions between the system and the user, in any case through rethinking the system architecture.

Building systems that embody ethical principles by design may also permit acquiring a competitive advantage in the market, as predicted in the recent Gartner Top 10 Strategic Technology Trends for 2019.²³

Computer scientists alone cannot solve the scientific and technical challenges we have ahead. A multi-disciplinary effort is needed that calls for philosophers, sociologists, law specialists, and computer scientists working together.

Acknowledgments. The author is indebted to the multi-disciplinary team of the EXOSOUL@univaq project (<http://exosoul.disim.univaq.it>) for enlightening debates and joint work on digital ethics for autonomous systems. □

References

- ACM U.S. Public Policy Council. Statement on algorithmic transparency and accountability, 2018; <https://bit.ly/2j4IJEV>.
- Larus, J. et al. When Computers Decide: European Recommendations On Machine-Learned Automated Decision Making, 2018; <https://dl.acm.org/citation.cfm?id=3185595>.
- EDPS. Opinion 4/2015: Towards a new digital ethics—data, dignity and technology; https://edps.europa.eu/sites/edp/files/publication/15-09-11_data_ethics_en.pdf.
- EDPS. Leading by example, The EDPS Strategy 2015–2019; <https://bit.ly/2Mpegij>
- European Group on Ethics in Science and New Technologies. Statement on artificial intelligence, robotics and 'autonomous' systems; https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf
- Burgess, J.P., Floridi, L., Pols, A. and van den Hoven, J. Towards a digital ethics—EDPS ethics advisory group; https://edps.europa.eu/sites/edp/files/publication/18-01-25_eag_report_en.pdf
- Floridi, L. Soft ethics and the governance of the digital. *Philosophy & Technology* 31, 1 (Mar. 2018), 1–8.
- The DECODE project; <https://decodeproject.eu>.
- Contissa, G., Lagioia, F. and Sartor, G. The ethical knob: Ethically customisable automated vehicles and the law. *AI and Law* 25, 3 (2017), 365–378.
- Gogoll, J. and Müller, J.F. Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics* 23, 3 (June 2017), 681–700.
- Ethics Commission Automated and Connected Driving. Appointed by the German Federal Minister of Transport and Digital Infrastructure, June 2017 Report; <https://bit.ly/2xx18DZ>
- Cath, C. et al. editors. Governing artificial intelligence: ethical, legal, and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Royal Society, Nov. 2018.
- Declaration on Ethics and Data Protection in Artificial Intelligence at the 40th Intern. Conference of Data Protection and Privacy Commissioners, Oct. 2018; <https://bit.ly/2Cz31AG>.
- AI Now Institute, New York University, 2017 Annual Report; https://ainowinstitute.org/AI_Now_2017_Report.pdf
- AI Now Institute, New York University, 2018 Annual Report; https://ainowinstitute.org/AI_Now_2018_Report.pdf
- Awad E. et al. The Moral Machine experiment. *Nature* 563 (Oct. 2018), 59–64.
- Li, T. China's influence on digital privacy could be global; <https://wapo.st/2TffDE0>
- Vardi, M. Are we having an ethical crisis in computing? *Commun. ACM* 62, 1 (Jan. 2019), 7; <https://cacm.acm.org/magazines/2019/1/233511-are-we-having-an-ethical-crisis-in-computing/fulltext>
- Nemitz, P. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical And Engineering Sciences*. Royal Society, Nov. 2018.
- Wakabayashi, D. California passes sweeping law to protect online privacy. *New York Times* (June 28, 2018); <https://nyti.ms/2tGjaAf>.
- The European Commission's High-Level Expert Group on Artificial intelligence, Draft Ethics guidelines for trustworthy AI; https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56433
- Artificial Intelligence: A European Perspective. European Commission Joint Research Centre, Dec. 2018; <https://ec.europa.eu/jrc/en/artificial-intelligence-european-perspective>
- Gartner Top 10 Strategic Technology Trends for 2019; <https://gtnr.it/2CJJYGp>

Paola Inverardi is a professor in the Department of Information Engineering Computer Science and Mathematics at the University of L'Aquila, L'Aquila, Italy.

BY THOMAS SKORDAS

Toward a European Exascale Ecosystem: The EuroHPC Joint Undertaking

HIGH-PERFORMANCE COMPUTING (HPC, also known as supercomputing) is indispensable in the new global data economy. The dramatic increase in the volume and variety of big data is creating new possibilities for sharing knowledge, carrying out scientific research, doing business, and developing public policies.

In almost every scientific discipline, researchers are using complex HPC modeling and simulation techniques and big data analytics approaches to answer fundamental questions and make new discoveries and breakthroughs. Representative examples include: in biology and the life sciences, understanding the dynamics of biomolecules and proteins and genome sequencing and genome assembly; in materials science and pharmaceutics,



designing new materials and discovering new drugs; and in cosmology and astrophysics, understanding dark matter. HPC also provides solutions to complex problems (such as understanding the organization and functioning of the human brain, developing new treatments based on personalized medicine, and predicting and managing the effects of climate change).

In industry, HPC paves the way for new businesses and innovative applications in high-added-value areas (such as manufacturing and engineering, particularly automotive and aeronautical engineering, oil and gas exploration, bioengineering and mo-



lecular chemistry, agri-food and precision agriculture, and developing and managing renewable and clean energies). In all these areas, HPC helps reinforce industrial innovation capabilities, particularly in small and medium-size enterprises (SMEs).

HPC is also becoming a key tool for public decision making in an increasing number of areas (such as in cybersecurity and defense, including developing efficient encryption technologies, and understanding and responding to cyberattacks and cyberwars, in nuclear simulations, in the fight against terrorism and crime, and in understanding and managing natural hazards and biological risks, in-

cluding earthquakes, flooding, failure of dams or power plants, and health pandemics).

HPC in Europe

In combination with artificial intelligence, HPC is set to become the engine to power the new global digital economy, where to out-compute is to out-compete. In this context, European supercomputing infrastructures represent a strategic resource for understanding and responding to the increasing challenges European citizens will face in the years to come, as well as for the future of European industry, SMEs, and the creation of new jobs. They are also key to ensuring European

scientists reap the full benefit of data-driven science.

However, today the European Union (EU) is not investing enough in HPC infrastructures and technologies to match its economic and knowledge potential, showing an annual funding gap of more than €500 million compared with the U.S. and China.

Europe needs an integrated leading-class HPC and data infrastructure with exascale computing performance that can compete worldwide. This infrastructure is absolutely necessary to ensure data produced in Europe by academia, industry, and SMEs is processed with its own su-

HPC is set to become the engine to power the new global digital economy, where to out-compute is to out-compete.

percomputing and data capabilities. This would reduce Europe's dependence on facilities in third countries and encourage innovation to stay in Europe.

Europe's Strategic Objectives

One of the priorities of the European Commission is to place Europe in the first three supercomputing powers of the world.^a To achieve this ambition, in October 2018 the EU, together with 24 EU member states and Norway, established the European High Performance Computing Joint Undertaking (EuroHPC JU),^b a public-private partnership between the EU, the 25 participating countries, and two European HPC and big data industrial associations—ETP4HPC^c and BDVA.^d

The EuroHPC JU is a legal and funding entity that makes it possible to combine EU and national funding with private resources. Its mission will be to develop, deploy, extend, and maintain in the EU an integrated world-class supercomputing and data infrastructure and support a highly competitive and innovative HPC ecosystem. EuroHPC will enable the EU and the participating countries to coordinate their supercomputing research agendas and investments and pool national and EU resources to close the funding gap with global competitors. The overall goals of the EuroHPC JU will be:

Acquire world-class supercomputers. To acquire world-class supercomputers available to European users from academia, industry, SMEs, and the public sector, including two pre-exascale systems by 2020, two exascale systems by 2022/2023 (at least one with European technology), and post-exascale supercomputers;

Maintain EU leadership. To maintain the EU's leadership in scientific and industrial applications, including HPC Competence Centres facilitating access to the HPC ecosystem, particularly for SMEs, and developing advanced digital skills; and

Independent and competitive HPC technology. To secure an independent

and competitive HPC technology supply for the EU, including future computing technologies (such as quantum).

Beside ETP4HPC and BDVA associations, other key European stakeholders in HPC and data infrastructures (such as PRACE^e and GEANT^f) are also expected to collaborate closely with the EuroHPC JU to establish a leading-class HPC, data, and communication infrastructure in Europe.

The EuroHPC JU will also establish a wide consultation process to inform and involve interested European stakeholders who can contribute to the realization of the EuroHPC strategy. In particular, the contribution of the European computer science and software communities and the participation of large scientific and industrial HPC user communities will be fundamental to realizing the strategic objectives of European world leadership in HPC applications and development of a competitive HPC technology-supply industry.

EuroHPC JU Ramp-Up Phase, 2019–2020

The EuroHPC JU will need substantial investments in order to achieve its goals. In the period 2019 and 2020, the EuroHPC JU will invest €1.4 billion, of which €1 billion will be public and €0.4 billion will be private investments. The EuroHPC JU will acquire and install two top-five pre-exascale machines and several mid-range supercomputers by 2020 and support activities across the full European HPC ecosystem, including:

Develop a European microprocessor and European exascale systems. The European Processor Initiative (EPI)^g partnership was recently established to implement a technology roadmap

e <http://www.prace-ri.eu/>

f <https://www.geant.org/>

g EPI is coordinated by ATOS/Bull and brings together 23 partners from 10 European countries, gathering experts from the HPC research community, the major European supercomputing centers, and the computing and silicon industry, as well as potential scientific and industrial users with an additional specific emphasis on automotive applications; <https://ec.europa.eu/digital-single-market/en/news/european-processor-initiative-consortium-develop-europes-microprocessors-future-supercomputers>

for future European low-power microprocessors for extreme-scale computing, big data, and emerging applications with a specific focus on exascale HPC and automotive computing for autonomous driving. Taking a co-design approach, EPI will design and develop the first European HPC systems on a chip and accelerators. Both elements will be implemented and validated in a prototype system that will become the basis for a full exascale supercomputer based on European technology.

Develop exascale software and applications. Exascale software and applications and their integration in extreme-scale prototypes will help ensure EU leadership in the application of HPC for scientific, industrial, and societal challenges. These activities will support development, optimization (including re-design), and scaling-up of HPC application codes toward exascale computing, as in HPC Centres of Excellence^h (CoEs). CoEs are also inherently committed to co-designing activities to ensure future HPC architectures are well suited for applications and their users, providing them with a high-performance, scalable application base.

Widen use of HPC and address the HPC-related skills gap. Widening the use of HPC and addressing the HPC-related skills gap will increase knowledge and human capital and boost HPC capabilities, including through creation of national HPC Competence Centresⁱ and their networking and coordination across the EU. These competence centers can gather the necessary resources and expertise to provide a single local entry point for customized HPC services, ranging from, for example, highly specialized scientific and technical HPC users to

SMEs with little or no experience in this domain.

EuroHPC JU in the Next EU Budgetary Period, 2021–2027

This is only the beginning, since in the EU's next Multiannual Financial Framework, covering the period 2021 to 2027, the aim is to continue supporting the EuroHPC JU via two different programs recently proposed by the European Commission—Horizon Europe^j and Digital Europe.^k

Horizon Europe is the EU's next Framework Program for Research and Innovation, the continuation of Horizon 2020,^l supporting the HPC research and innovation agenda and addressing exascale and post-exascale technologies.

The European Commission is also proposing to support the EuroHPC JU via the Digital Europe program with an additional €2.7 billion. This will cover acquisition of exascale supercomputers (at least one with European technology in 2022/2023) and post-exascale systems (around 2027), including integration and deployment of the first hybrid HPC/quantum infrastructure in Europe; and actions to develop advanced HPC skills and further facilitate access to industry, academia, and public administrations to the HPC ecosystem, and more. The program will also exploit the synergies between HPC and other digital priorities, including artificial intelligence, cybersecurity, digitizing public-sector services, and digital skills.

Conclusion

The EuroHPC JU is an ambitious initiative that will enable European countries to coordinate with the European Union their supercomputing strategies and investments. We thus need to reduce the fragmentation of HPC investments across Europe and align strategies and investments that are key for European innovation and competitiveness. We need to secure access to

^h <https://ec.europa.eu/digital-single-market/en/news/eu-funded-hpc-research-projects-and-centres-excellence-nutshell>

ⁱ A national HPC Competence Centre is a legal entity established in a participating state that is a member state associated with the national supercomputing center of that member state, providing users from industry, including SMEs, academia, and public administrations, with access on demand to supercomputers and to the latest HPC technologies, tools, applications, and services, and offering expertise, skills, training, networking, and outreach.

^j https://ec.europa.eu/info/designing-next-research-and-innovation-framework-programme/what-shapes-next-framework-programme_en

^k http://europa.eu/rapid/press-release_IP-18-4043_en.htm

^l <https://ec.europa.eu/programmes/horizon2020/en/>

world-class data and supercomputing facilities across Europe, ensuring development in Europe of an integrated exascale supercomputing capability accessible throughout Europe and covering the whole value chain, from technology components to systems and machines to applications and skills. This will avoid long-term negative effects on Europe's data sovereignty and scientific and industrial leadership and on Europe's place in the digital economy at large.

Today, the EU has put HPC very high in the political agenda, and we are confident this will remain the case for a long time. We are witnessing an exceptional surge of positive dynamics contributing to the success of the EuroHPC JU, including wide political support at both the national and EU levels, very strong support from public and private stakeholders, and the convergence of HPC with other critical disciplines for the data economy (such as big data and artificial intelligence). This is a unique opportunity for Europe to reap the benefits of mastering these converging technologies for our future. Only by joining forces can we mobilize substantial European and national efforts, both public and private, to place Europe in a leading position in the global digital economy.

Disclaimer. The views expressed in this article are the sole responsibility of the author and in no way represent the view of the European Commission and its services. □

Thomas Skordas is Director of Digital Excellence and Science Infrastructure, DG Communications Networks, Content and Technology, European Commission, Brussels, Belgium.

DOI:10.1145/3312569

BY STEFFEN STAAB, SUSAN HALFORD,
AND WENDY HALL

Web Science in Europe: Beyond Boundaries

AS WE FINALIZE this article November 11, 2018, and consider current and future directions for computing in Europe and across the globe, we remember the end of World War I exactly 100 years ago: the end to a war of atrocities at a scale previously unseen and the culmination of a series of events that European nations had allowed themselves to ‘sleepwalk’ into, with little thought for the consequences.¹⁰

When this article appears in spring 2019, we will remember the first proposal for a new global information sharing system written by Tim Berners-Lee 30 years ago at CERN,⁴ the European organization for nuclear research. This proposal marked the beginning of the World Wide Web, which now pervades every facet of modern life for over four billion users. However, the Web 30 years on, is not the land of free information and discussion, or an egalitarian space that supports the interests of all, as originally imagined.⁴ Rather, egotisms, nationalisms, and fundamentalisms freewheel on a landscape that is

increasingly dominated by powerful corporate actors, often silencing other voices, including democratically elected representatives.

For seven decades Europe has been a political and social project, seeking to integrate what has been divisive historically and to make citizens more equal. While the proponents of the Web were driven by similar values, there is now increasing concern in Europe—and beyond—that the Web has become a vehicle of disintegration, polarization, and exploitation. What is more, since the Web operates at a global scale, beyond nation-states and with little formal regulation, we lack both the understanding and the means to avoid sleepwalking into another catastrophe.

Web Science seeks to investigate, analyze, and intervene in the Web from a sociotechnical perspective, integrating our understanding of the mathematical properties, engineering principles, and the social processes that shape its past, present, and future.⁷ Over the past 10 years, Web Science has made remarkable progress, providing the building blocks to face the challenges described here. And yet there is more do to. In this article, we offer a more detailed definition of Web Science and outline its achievements to date. We consider how Web Science frames and addresses key sociotechnical challenges facing the Web now and for the near future, emphasizing the importance of this as new artificial intelligences start to shape the Web (and Web Science) in significant new directions. Arising from this, we outline some of the practical strategies Web Science is developing to integrate knowledge across disciplinary boundaries and build collaboration with Web stakeholders. Web Science equips us to understand the past and present of the Web and the skills and tools to shape a positive future.

What Is Web Science?

Web Science in Europe begins from the premise that the Web is both technical



and social. From this perspective, it is so difficult to disentangle the social from the technical that we describe the Web as ‘sociotechnical.’ The Web has been built on layers of communication at different levels of abstraction, from physical link layers (such as Ethernet) over Internet and transport layers (such as TCP/IP). It started as a Web of Documents (HTML), which served as the nucleus that other Webs would piggyback on: a Web of Data (RDF, SPAR-

QL), a Web of Services (REST, JSON), a Web of Things.^a

All these layers are defined by underlying technical standards and are the result of sophisticated engineering. And they are also deeply social, in two key ways. First, they have been developed in particular social contexts, with social goals in mind. For example, CERN was established to ensure

a European nuclear capacity after the devastation of the research infrastructure in World War II.¹³ Similarly, the original intentions for the Web were to allow physicists to share data across teams underpinned by an intellectual commitment that information ‘wants to be free.’⁸ Second, the Web merely offered a set of opportunities for humans to develop and populate information constructs and link with each other. Over time we have seen multiple and

a <https://www.w3.org/WoT/>

How do we engage the public in meaningful dialogue and decision making about the future of the Web?

competing rationalities drive the take-up of these opportunities. For example, information sharing and community building dominated academic and countercultural use in the early days. As new users began to embrace the opportunities on offer—for government and commerce in particular—content began to change. More than this, new users began to shape Web technologies—for example enabling user-generated content, video streaming, and secure online payments—in ways that, in turn, opened up new possibilities both positive, and less so.

The Web has changed the world and the world has changed the Web. And this is only set to continue, as the platform economy, the Internet of Things and new artificial intelligences offer new opportunities and shape the Web into the future.

For the past decade, Web Science has been building the interdisciplinary expertise to face the challenges and realize the value of this rapidly growing and diversifying Web. This task transcends the work of any single academic discipline.⁷ While our universities continue—overwhelmingly—to be organized in silos established in the 20th century, or much earlier, the Web demands expertise from computer science, sociology, business, mathematics, law, economics, politics, psychology engineering, geography, and more. Web Science exists to integrate knowledge and expertise from across fields, integrating this into systematic, robust, and reliable research that provides an action base for the future of the Web.

Evidence of our endeavors includes the networks of Web science labs, a number of undergraduate and postgraduate educational programs across Europe, summer schools on Web Science, and an ACM conference series.^b We have understood how we may target to build ‘objective’ technology, yet end up with social stereotypes we wanted to avoid.² We have learned about the social and the technical processes that are needed to provide open data for the social good,^c the methodological and epistemological challeng-

es of using new forms of digital data and computational methods for social research,^{15,16} and Web Science has progressed Social Machines that let us collaborate, yet work independently in distributed fashion.²³

And yet there is much more to do. As a topical and critical example, we need to understand how the Web influences our democracies. Democracy builds on pillars like the representation of all, the rule of law, publicity and quality of information, temporality of decisions, and autonomy of individuals. The Web affects these pillars: online intimidation may threaten individuals and, silence them. Groups may organize online to ignore the law. Misinformation in echo chambers lowers the publicity and quality of information. In light of too much online transparency, compromises—which are vital in democracy—become infeasible. And, autonomy may be jeopardized by intrusion into private spheres. For all that, the Web continues to offer positive opportunities—voice to the otherwise silenced, connections between fragmented populations, mobilization of those who lack other means or are repressed—it is clear that these opportunities have come at a cost and—more broadly—that we may need to reconsider the pillars of democracy in digital society. These questions make Web Science more important now than ever. While Europe strives to respond to them in EU projects,^d and various national endeavors thrive (for example, the Alan Turing Institute in the U.K. and, the German Internet Institute) we have only begun to face the challenges.

The Sociotechnical Challenges

There is nothing inevitable about the future of the Web. Its history to date has been made at the intersection of technical innovation and everyday practice with wider social processes and power relations, defying any prediction of fixed or finished outcomes. While this poses profound challenges—we cannot simply engineer the Web into a preferred state—we must develop integrated and in-depth sociotechnical understandings of the Web if we are to influence its future direction.

b <http://webscience.org> on labs, conference, educational programs and summer schools.

c <https://theodi.org/>

d For example, <http://coinform.eu/>

Here, we describe two key developments that characterize the opportunities and challenges we face.

Datafication refers to the development that our everyday activities are traced digitally at unprecedented scale and accuracy for commercialization and exploitation in a data economy. Datafication raises questions about how this situation can or should be managed and what might result out of its pervasiveness. The processes of datafication, their consequences and how we live with these are both social and technical. From the beginning, the question of what data is created depends both on human activities and technical devices.

How this data is used depends on configurations of ownership, markets, state authority, and citizens' rights as well as the technical affordances for circulation through technical infrastructures and the computational possibilities for analysis. To even describe the processes of datafication demands expertise of the highest level from computer science, law, political science, sociology, and more. To consider if and how society might respond to this new landscape likewise. What are the opportunities to flip data ownership from the big tech companies to the individuals whose data fuels the data economy? Engineering solutions, as developed in the SoLiD^e project, may be part of the response, but how can we be sure that people even want let alone will have the capacity to use these solutions? What new challenges might these solutions pose? How would this impact on the underlying business model for the Web?

The digital divide. Web access continues to rise rapidly but over three billion people worldwide have no access, and 1:8 of the European population does not use the Web regularly.^f We should avoid normative claims that the Web is 'good' for everyone, we know now that this is not the case, yet at the same time this should be a matter of choice not constraint. Further, beyond the question of access alone, we see an increasing divide between those highly skilled users who are able to derive the

greatest benefit and those less skilled who are less knowledgeable about privacy risks, less able to protect their security and may derive less economic benefit from the opportunities available online.¹⁷ So long as people are unaware of the technical mechanisms and social uses of datafication or the potential effects of this on their lives and life chances they will not be able to make effective choices about how to use the Web or join the public debate about the future of the Web. Web Science calls for new approaches to digital literacy, beyond the use of Web tools and beyond the extension of coding skills to schools (important as both these are) to build understanding of the Web as a sociotechnical system and drive toward greater empowerment of Web citizens. It engages, for example, through the Web We Want campaign, #fortheweb, and educational interventions.¹⁸

Both these examples are linked to wider practical, political, and philosophical questions. What are the checks and balances with regard to openness and privacy? What forms of transparency and accountability are appropriate and achievable, to balance individual privacy, fairness across social groups and a viable business model for the future of the Web? How do we engage the public in meaningful dialogue and decision making about the future of the Web?

Next, we investigate another most prominent sociotechnical challenge in more detail that today is most often characterized as a technical challenge alone, whereas it is deeply entrenched into the way that we as individuals or as society interact with each other and with the artifacts we create.

Web and Artificial Intelligence

The Web and its infrastructure has become interwoven not only with documents, but also with data, services, things—and artificial intelligences. Initially, the Web was a field of application for artificial intelligence. Knowledge-based systems and machine learning were used to provide intelligent access to information on the Web, to enhance search, to facilitate browsing or to negotiate in electronics market. In hindsight, this may be considered to have been a very useful,

but a shallow, piecemeal interaction between Web and AI.

Yet since the end of its first decade, there was a vision to build a Web that was intelligent in itself, that included agents that would assist its users.⁶ As this objective was beyond reach then, the Semantic Web community increasingly refocused on what became a proverb that data with *a little semantics goes a long way*. When researchers started to properly understand and use the social motivation of Web developers and Web content managers, some European researchers developed what now has become the two most popular Semantic Web applications, Wikidata²⁷ and Schema.org.⁸ At the same time Web Science was coined as a field that would address the systematic understanding of these socio-technical interactions between Web and humans.⁷

At the end of the second decade of the Web, artificial intelligence took several major turns. Big data, which frequently came from the Web directly or from crowdsourcing on the Web, became the foundation for human-like performance on some tasks such as image annotations.¹⁹ At the same time chatbots and virtual assistants have been developed and are now widely found on our PCs, smartphones, and in our homes. The latest developments let these virtual assistants acquire their knowledge from the Web, from archived dialogues,¹² or from live interaction.

Microsoft researchers were pushing the edge and put their AI chatbot "Tay" online to interact with and learn from human encounters. Humans quickly taught it to go <*from "humans are super cool" to full nazi in <24hrs>*>.^h While there was a wide discussion that the technology was inadequate, there seemed to have been little understanding that it was the social context and the social processes that determined the fate of Tay. While in the initial Semantic Web, the lack of such understanding led to a simple, but not very problematic non-adoption, in the case of Tay being an active agent the lack of

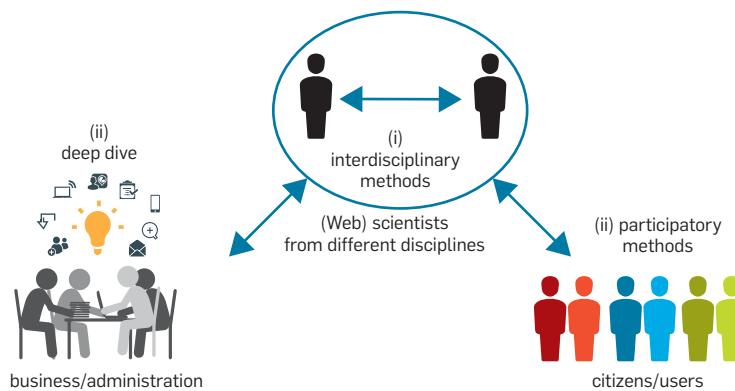
^g Schema.org was an agreement of several search companies modeled after the preceding Yahoo! SearchMonkey system.²¹

^h <https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

^e <https://solid.mit.edu/>

^f <https://www.statista.com/topics/3853/internet-usage-in-europe/ref>.

Web Science methods must remain incomplete, if lacking interaction between scientists (i), or if not involving all of the Web's stakeholders.



insight led to malbehavior.

The Web as a social medium, whether considering past contributions or ongoing interaction, is prone to misguide artificial intelligences. Indeed, the question arises, what the social values are that an AI on the Web should embed and how this should be realized? Efforts to censor the successor of Tay by ruling out topics like religion and politics hamper the chatbot leaving it socially awkward.²⁵ Notions of social biases² and data representativeness are interwoven, but who decides whether or when the answers are ‘right’? Several researcher communities (for example, Semantic Web, Computer-Human Interaction) and institutions have decided to actively tackle some underlying problems, for example, addressing the underrepresentation of women on Wikipedia by Edit-a-thons.

Finally, in two decades the Web has produced a range of highly valuable companies that did not play a major role before, or were not even founded when the Web started. Many of them benefit from first-mover and network effects that are difficult if not impossible to imitate by new companies. Will few big AI companies use their intellectual and computational power to rule the world using AI in the future? Or can society draw close, organizing the many and by sharing the necessary data and computation power bring AI to everyone’s fingertips? The CommonVoice projectⁱ is certainly a project of developing AI on the Web in a direc-

tion that benefits more than a few of the already wealthy.

Extending Web Science

Web Science in Europe has begun the task of building up a body of knowledge to address these challenges. (Further information on the Web Science conference, educational programs and summer schools can found at <http://www.webscience.org/>.) Yet we have more work to do in extending Web Science, both within and beyond the academy. We classify the challenges by considering the interaction between various stakeholders involved, as illustrated in the accompanying figure.

Interdisciplinary methods. To the present day, the vast majority of Web research is disciplinary. Web Science in Europe has been at the forefront of developing interdisciplinary approaches to describing, analyzing and intervening in the Web. Our experience over the past decade shows that working across disciplines brings a depth of analysis and level of confidence in research outcomes that is much needed to address the very real challenges facing the Web—and society—as we move forward into the 21st century. Our experience also allows us to see where we can and should extend Web Science research through the novel application and development of research collaboration.

We are the first to recognize that this is challenging. Academic disciplines work with different objectives and have crafted a range of epistemologies, methodologies, and methods that have distinct professional standards. This is

particularly noticeable across the computational and social sciences, where there are some profound differences in what counts as knowledge, science, and method. This is evident in the majority of—otherwise exciting—conferences between the social sciences and computer science, which tend to start from one ‘side’ or the other, and to privilege that body of knowledge, rather than opening it to revision and reconstruction through engagement from beyond.

Web Science has made the case for interdisciplinarity at a high level but transcending these established knowledge frameworks to build new understandings is difficult, demanding creativity, risk taking, and generosity.

One of many examples we may envision is the use of **interdisciplinary visual data analytics**. Web data offer remarkable potential to analyze the things that people say and do, in real time, over time, rather than the things that they say they do when asked using conventional methods, for example, interviews and surveys.^{21,26} However, integrating understanding of the data and the computational methods required to interrogate this data with the domain-specific expertise required to address specific questions is challenging.¹⁷ Furthermore, developing robust methodological understanding of the data and the effects of applying particular computational methods to this data is, as yet, in its infancy. While the visualization community in computer science harbors a wealth of techniques and tools to interactively explore data and find patterns, joint research work that would give Web scientists the means to ‘interview’ Web data and trace the impact of computational methods on results are lacking. Visualizations approachable and understandable across Web scientist subcommunities might become ‘boundary objects’ enabling different forms of expertise focus on the same phenomenon.

Another example is **participatory methods**. Much has been said about the ignorance of researchers about what the broad public wants, as well as about the ignorance of the broad public about what the scientists deliver. Let us consider the example of privacy protection. While the public’s insight

ⁱ <https://voice.mozilla.org/>

into understanding implications of privacy issues may have been limited, one might have acknowledged that the public's attitude toward privacy protection did not only stem from lack of knowledge, but also from some nuanced degrees of willingness to share personal information. Such an ambiguous situation calls out for a two-way, participatory dialogue. Not content with only researching 'on' users, Web Science is committed to ensuring that the full range of voices is heard as we build our understanding of the Web and shape its future. Web Science seeks creative ways to build public understanding of the public about the threats, but also take on board, appreciate, and remark upon the personal values and attitudes of people. For instance, moral machines are one example where this is done now.³ We are committed to developing participatory methods that allow us to build insight to diverse perspectives and to build dialogues between these. These methods may include: citizen science—where non-experts are included in a variety of research projects, for example, to study local communities^j or to contribute subjective, possibly diverging, point of views;¹ online methods for deliberation; organizing face-to-face citizens' assemblies; and the use of AI techniques (for example, for enhancing knowledge and understanding of the Web and extending dialogue). It is a priority for Web Science that we observe these processes in action to inform continuous improvement in public engagement, for the benefit of policy making and, more widely, the engineering of the Web.

A final example concerns how we observe **the observers**. Powerful corporate or governmental actors may determine the fate of Web users observing what we do²² and suggesting what we might do (or not), for instance, which accommodation to select, which job to apply to, or which person to befriend. Therefore, understanding what these actors do by tracking their activity and evaluating their algorithms has become an important activity. Researchers and NGOs like Algorithmwatch^k pursue these tasks asking for data do-

nations or crowdsourcing for getting insight into potentially discriminating or exploitative behavior. In other realms of life, corporate actors need to prove their carefulness by admitting to oversight of governmental agencies. In the Web we still lack such regulations, but the more that such actors become gatekeepers to our life, the less we can just rely on corporate slogans like "Don't be evil" (originally used in Google's corporate code of conduct).

Conclusion

The Web has grown from an idea in 1989 to become the largest sociotechnical assemblage in human history in a little under 30 years. It is implicated in the lives, livelihoods, and life chances of over half the world's population already and connecting many more every day. While Europe embraces the Web and its opportunities for integration—perhaps more than other parts of the world—it discusses its risks of division. Rather than dystopian, and most likely false, predictions, what it needs is a scientific approach to understanding how the Web works and how it affects society. Web Science has been devised as a field to tackle these questions and we have highlighted a few aspects of where and how Web Science should proceed. In particular, computer science must look beyond its pasture and embrace the methodological experience and diversity by a broad set of fields—more than it has done until now. Funding and academic institutions need to welcome and reward such undertaking or it will not succeed.

Acknowledgment. This article benefited immensely from discussions we had with all the other participants at the Dagstuhl^l seminar on "10 Years of Web Science: Closing The Loop." In particular, we want to thank Bettina Berendt, Fabian Gandon, Katharina Kinder-Kurlanda, and Eirini Ntoutsi. **C**

¹ <https://www.dagstuhl.de/en/program/calendar/semp/?semnr=18262>

References

- 1 Arroyo, L. and Welty, C. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (Jan. 2015), 15–24.
- 2 Baeza-Yates, R.A. Bias on the Web. *Commun. ACM* 61, 6 (June 2018), 54–61.
- 3 Bello, P. and Bringsjord, S. On how to build a moral machine. *Topoi* 32, 2 (Oct 2013), 1572–8749.
- 4 Berners-Lee, T. *Information Management: A Proposal*. Technical Report, CERN (Mar. 1989, May 1990); <http://cds.cern.ch/record/369245/files/dd-89-001.pdf>
- 5 Berners-Lee, T. *Weaving the Web*. Harper, New York, 2000.
- 6 Berners-Lee, T., Hendler, J. and Lassila, O. The Semantic Web. *Scientific American* 284, 5 (May 2001), 34–43.
- 7 Berners-Lee, T. et al. Creating a science of the Web. *Science* 313.5788 (2006), 769–771.
- 8 Brand, S. *The Media Lab: Inventing the Future at MIT*. V. Viking Penguin, 1987.
- 9 Cunningham, J. *Digital Exile: How I Got Banned for Life from AirBnB*. <https://medium.com/@jacksoncunningham/digital-exile-how-i-got-banned-for-life-from-airbnb-615434c6eeba>
- 10 Clark, C. *The Sleepwalkers: How Europe Went to War in 1914*. Penguin, 2013.
- 11 Day, M. *Teaching the Web: Moving Towards Principles for Web Education*. Ph.D. dissertation, University of Southampton, 2019.
- 12 Gao, J., Galley, M., and Lihong, L. *Neural Approaches to Conversational AI*. (2018); CoRR abs/1809.08267
- 13 Gillies, J. and Cailliau, R. *How the Web Was Born*. Oxford University Press, Oxford, 2000.
- 14 Halford, S. and Savage, M. Reconceptualising digital inequality. *Information, Communication, and Society* 13, 9 (July 2010), 937–955.
- 15 Halford, S. *Digital Futures? Sociological challenges and opportunities in the emergent Semantic Web*. *Sociology* 47, (Jan. 2012), 173–189.
- 16 Halford, S. et al. Understanding the production and circulation of social media data: Towards methodological principles and praxis. *New Media and Society* (2017); <https://doi.org/10.1177/146144817748953>.
- 17 Halford, S. and Savage, M. Speaking sociologically with big data: Symphonic social science and the future for big data research. *Sociology* 51, 6 (June 2017), 1132–1148.
- 18 Hill, B.M. Almost Wikipedia: Eight early encyclopedia projects and the mechanisms of collective action. In *Essays on Volunteer Mobilization in Peer Production*. Ph.D. dissertation, Massachusetts Institute of Technology, 2013. https://mako.cc/academic/hill-almost_wikipedia-DRAFT.pdf.
- 19 Krizhevsky, A. et al. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (June 2017), 84–90.
- 20 Mika, P. and Tummarello, G. Web semantics in the clouds. *IEEE Intelligent Systems* 23, 5 (May 2008), 82–87.
- 21 Savage, M. and Burrows, R. The coming crisis of empirical sociology. *Sociology* 41, 5 (May 2008), 885–899.
- 22 Schelter, S. and Kunegis, J. On the ubiquity of Web tracking: Insights from a billion-page Web crawl. *J. Web Science* 4, 4 (Apr. 2018), 53–66.
- 23 Shadbolt, N.R. Towards a classification framework for social machines. *WWW (Companion Volume)* 2013, 905–912.
- 24 Simonite, T. When it comes to gorillas, Google photos remains blind. *Wired* (Jan 11, 2018); <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.
- 25 Stuart-Ulin, C.R. Microsoft's politically correct chatbot is even worse than its racist one. *Quartz* (July 31, 2018); <https://qz.com/1340990/microsofts-politically-correct-chat-bot-is-even-worse-than-its-racist-one/>.
- 26 Tinati, R. Big data: Methodological challenges and approaches for sociological analysis. *Sociology* 48, 4 (2014), 663–668.
- 27 Vrandecic, D. and Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* 57, 10 (Oct. 2014), 78–85.

Steffen Staab holds a chair for Web and computer science at the University of Southampton, U.K. and is a professor at the Universität Koblenz-Landau, Germany, heading its Institute for Web Science and Technologies (WeST).

Susan Halford is a professor of sociology at the University of Bristol, U.K.

Dame Wendy Hall is Regius Professor of Computer Science at the University of Southampton, U.K. and is the Executive Director of the Web Science Institute.

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

^j <https://bit.ly/2SF8O1w>

^k <https://algorithmwatch.org/en/>



The complex cacophony of intertwined systems.

BY PAT HELLAND

Identity by Any Other Name

“What’s in a name? That which we call a rose by any other name would smell as sweet.”

—William Shakespeare (*Romeo and Juliet*)

AS DISTRIBUTED SYSTEMS scale in size and heterogeneity, increasingly identifiers connect them. These may be called IDs, names, keys, numbers, URLs, file names, references, UPCs (Universal Product Codes), and many other terms. Frequently, these terms refer to immutable things. At other times, they refer to stuff that changes as time goes on. Identifiers are even used to represent the nature of the computation working across distrusting systems.

The fascinating thing about identifiers is that while they identify the same “thing” over time, that referenced thing may slide around in its meaning. Product descriptions, reviews, and inventory balance all change, while the product ID does not. Reservations, orders, and bookings all have identifiers that do not change, while

the stuff they identify may subtly change over time.

Identity and identifiers provide the immutable linkage. Both sides of this linkage may change, but they provide a semantic consistency needed by the business operation. No matter what you call it, identity is the glue that makes things stick and lubricates cooperative work.

This article is yet another thought experiment and rumination about the complex cacophony of intertwined systems.

The Need for Identity

For a long time, we worked behind the façade of a single centralized database. Attempting to talk to other computers was considered an “application problem” and not in the purview of the system. Data lived as values in cells in the relational database. Everything could be explained in simple abstractions, and life was good!

Then, we started splitting up centralized systems for scale and manageability. We also tried to get different systems that had been independently developed to work together. That created many challenges in understanding each other⁴ and ensuring predictable outcomes, especially for atomic transactions.

As time moved on, a number of usage patterns emerged that address the challenges of work across both homogeneous and heterogeneous boundaries. All of those patterns depend on connecting things with notions of identity. The identities involved frequently remain firm and intact over long periods of time.

Data on the outside vs. data on the inside. In 2005, I wrote a paper, “Data on the Outside versus Data on the Inside,”⁷ that explored what it means to have data not kept in the SQL database but rather kept in messages, files, documents, and other representations. It turns out that information not kept in databases emerges as immutable messages, files, values (à la key/values), or other representations. These are typically semi-structured in their representations, but *they always have some form of identifier*.



Scale, long-running, and heterogeneous. Systems are knit together by identity, too. As homogeneous solutions are designed for scale, shards, replicas, and caches are all based on some form of identity. Solutions respond to stimuli over time, using one or more representations of identity to figure out what work to restart or continue. Connecting independently created systems with their own private and distrusting implementations always uses shared identities and identifiers that are the crux of their cooperation.

Searching and learning. Many other parts of the computing landscape depend on identities. Searching assigns document IDs and then organizes indices of search terms associated with them. Machine learning binds attributes with identities. In many cases, a set of attributes becomes interesting and is then assigned an identity. The system repeatedly works to associate even more attributes to them. It's

when these attributes form patterns across the identities that the machine has learned something.

Identities: The new fulcrum. Computing patterns show our dependency on identities. We used to look only at relational databases but now we see pieces of computation and storage interconnected by identities. The data and computation connected by identities can swirl and shift around.

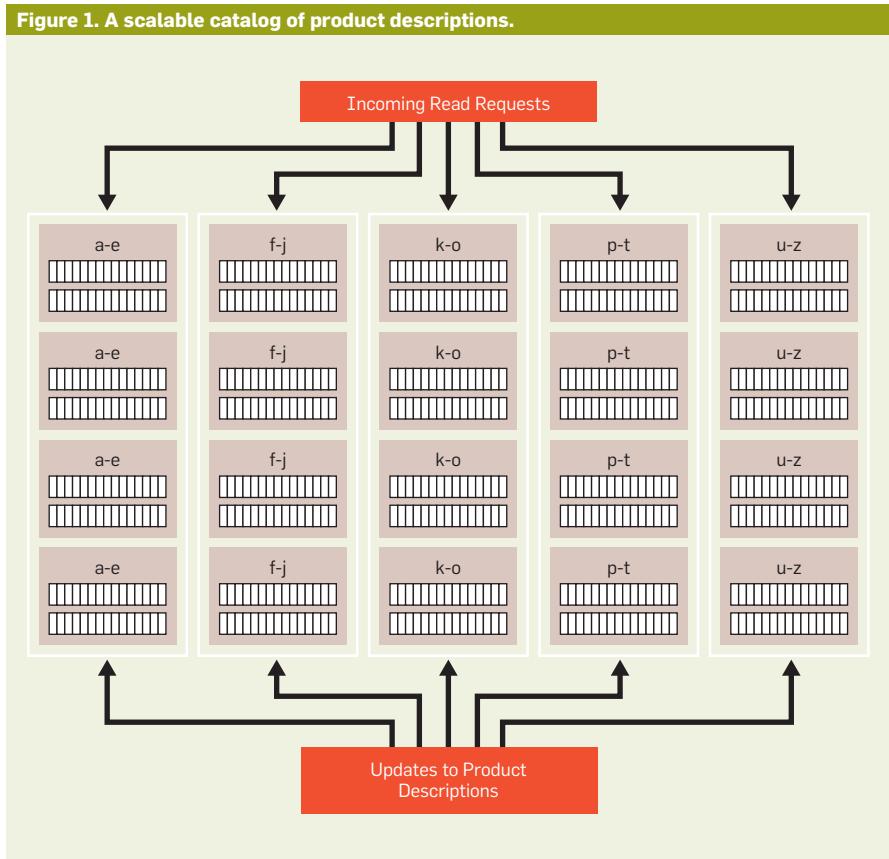
The identifiers connecting these pieces remain immutable while the stuff they identify spins and dances and evolves. Similarly, whatever is using the identity may be simply a mirage while the identifier used remains solid.

What's in a name? This article refers to *identities*. There are an astonishing number of synonyms for identity. All that really matters is that the identity is unique within the spatial and temporal bounds of its use. Name, key, pointer, file name, handle, check number, UPC, UUID (universally unique

identifier), ASIN (Amazon Standard Identification Number), part number, model number, SKU (stock keeping unit), and more are unique either globally or within the scope of their use. It is the immutable nature of each identifier *within the scope of its use* that allows it to be the interstitial glue that holds computation together.

Using Identity to Scale

Identity may be used to scale homogeneously and heterogeneously. This section examines a very complex example. E-commerce not only uses shopping carts and scalable product catalogs, but it may also derive product descriptions by combining the best information from some of its many merchants and manufacturers. Merchant SKUs or manufacturer part number may identify this information. In addition, inventory, pricing, and condition of offered goods all vary by merchant and are identified in a nonstandard way. Many

Figure 1. A scalable catalog of product descriptions.

different connected and disconnected identities weave through the complex multiccompany e-commerce.

Session-state and shopping carts. Each shopper gets their own shopping cart. This can be associated with an online account or with the Web session. Shoppers do not get multiple shopping carts during a single Web session. Furthermore, no one expects or wants the shopping cart to share state or consistent updates with other shopping carts.

The uniqueness of the shopping cart is provided by the shopping cart ID. There is some logic in the system to bind the session, either via user login or online session state, to a shopping cart ID. Based on that unique ID, the shopping cart contents are located.

The scalable key-value store. One common pattern in scalable solutions is the scalable key-value store. Take, for example, an e-commerce retail product catalog. The retailer has a whole bunch of products, each with a product identifier. The product description cache is sharded by the product ID. This supports scalable description data. Replicated shards support scalable read traffic. To add more product descriptions, add more shards. To support

more read traffic, add more replicas of the shards. See the scalable catalog of product descriptions indexed by the product ID in Figure 1. There is no requirement that the product catalog can update different products atomically. In fact, the product catalog cannot update all the cached entries for a single product atomically!

Identifying cached jittery versions. Updates to product descriptions distribute new versions to replicas over time. Hence, reads are jittery, and later reads may show earlier values. Product ID is the immutable glue that makes this work. Even if the read of the cache returns an old cached value, it is associated with the desired product ID and meets the business needs. In product catalogs and for many other uses, old values are fine.

Matching and deriving descriptions. In most large e-commerce sites, product descriptions come from data submitted by manufacturers, merchants, and other sources. To correlate these, it is necessary to normalize inputs, match descriptions from different sources, and then combine them to get the best information available. Inputs arrive with identifiers such as model number, UPC, and SKU, defined by the

third-party merchant selling through the large e-commerce site. There's no single identity before matching.

Normalizing cleans up the various inputs to try to have a consistent representation. If the color is Kelly green, forest green, olive, or chartreuse, should it be normalized to *green*? Normalization makes it easier to match various inputs to each other. It also loses some of the fidelity of the original input.

Matching attempts to find stuff that is the same. Is this product for sale from Merchant A the same as another product for sale from Merchant B? Each merchant has their own SKU as a personal unique identifier. How can they be correlated?

Slippery and sliding identifiers. Another challenge is that the merchants' SKUs are assigned and bound by the merchants. There's nothing to stop them from changing SKU 12345 from a pair of ruby slippers to a can of chocolate sauce. *When your partner business uses identifiers in a non-immutable way, you need to be on your toes.* I've heard tales of small merchants with 40 bins of stuff in their basement. The contents of SKU #23 corresponds to whatever product is kept in bin #23 at the time.

UPCs: The same but ... maybe different. Consider large retailers that consolidate many sellers' goods through the large retailer's platform. It is helpful if the merchants have the UPCs in the description of their item(s). UPCs make it much easier to match items from different merchants. Each of these 12-digit identifiers is for a *particular manufactured product*. The UPC works along with the EAN-13 (European Article Number 13) code, which is a bar code supporting scanners mostly for retail environments.

UPCs are *mostly* correct. Achieving consistency and equivalence of products with the same UPC is hard for both manufacturing and retail. Not everything has a UPC. Handcrafted items, for example, may not have UPCs. For a number of years, shoes were notorious for not having UPCs.

Books: ISBNs, paperback, used, and digital. What about books? The International Standard Book Number (ISBN) is a 13-digit (formerly 10-digit) number that uniquely identifies a particular *version and format* of a book.

What about reviews? Most reviews

are about the contents of the book, not the quality of the paperback's binding. Don't you want to have shared reviews for the e-book, paperback, and hardback editions? Typically, this is handled with yet another unique identifier used to represent all the different versions and formats. Similarly, many times the same online products share reviews when the color and unique identifier differ.

Products, SKUs, offers, inventory, and shippability. Online retail is an ocean of unique IDs, all weaving across different systems, concepts, and cooperating companies. Merchants will describe *their perspective* of goods for sale as their SKUs. Matching and correlating these goods into products from the perspective of the e-commerce site is a major endeavor in data science and machine learning. When done, the correlation is kept to facilitate working across the merchant and e-commerce site. Of course, the merchant is free to label a completely different product with the same SKU tomorrow; the e-commerce site must adapt.

The identifiers for products will reference the product catalog. The contents of the product catalog will evolve and be cached for efficient scalable reads. When accessing the cache, it may race with updates to the cache, and later reads may return earlier versions of the product description. It doesn't matter because either version is OK. The product catalog does not need transactional consistency.

Next, an offer to buy from a merchant is presented. Do you want a new or used product? What condition is it in, and what's the reputation of the seller? These offers are correlated to the product, the shopping cart, the inventory for the specific offer, the price, the shipping commitment, and the details of how it will be shipped. Of course, this needs to be tied to the payment.

Each of these relationships across internal and external systems is knit together using various related identities. Figure 2 shows a very small subset of these interactions and how identifiers knit them together. Oh, yeah; the e-commerce retailer hopes the merchant has not recycled the SKU when an order is placed. Attaching the product description to the SKU usually avoids confusion.

Using Identity to Search

Let's consider Web search as we have all seen it in Yahoo!, Google, and Bing. Not surprisingly, searches are accomplished by assigning unique IDs to each of the documents in the Web.

Document IDs, URLs, and search terms. As these huge Web crawlers traverse the URLs they find to locate documents, they remember the URL for each document. These URLs form unique IDs. It's common to bind the URL to another unique document ID that's shorter.

As the document is crawled, the word sequences are extracted for indexing. These word sequences (known as N-grams) correspond to the search terms entered into the Web search application.

N-grams are sharded into a large number of partitions. As multiple search terms enter a search, the shards that may hold those terms are queried. This returns sets of document IDs from many shards. By comparing the results looking for document IDs in common across the search terms, a resulting collection of document IDs can be returned.

While this is vastly and grossly oversimplified, the main point is that search is all about identities.

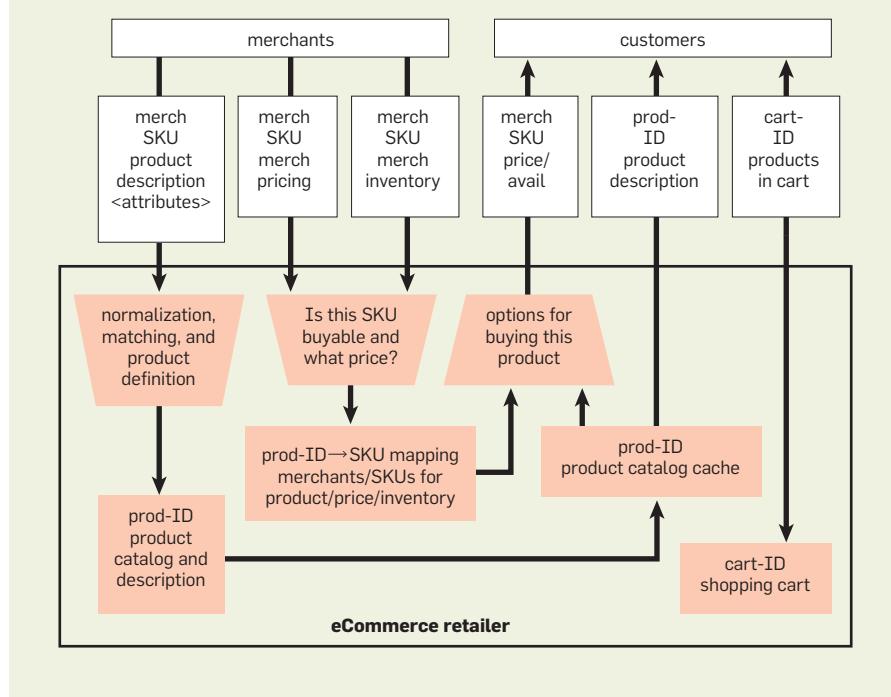
Searching an object-relational app.

Object-relational systems typically have application objects layered on top of underlying relational systems. Some object-relational systems offer search features that find the identities of objects based on their contents and the N-grams within them. This mechanism depends on the object identities captured by the search system and correlated to the objects. While these identities may not be explicitly understood by the underlying SQL database, they are understood by the object-relational system and the search engine layered on top.

Search means finding identities. Search today typically means a system that finds identities of documents, objects, or other things. It is the correlation of the N-grams extracted from these things to the identities that provides search results. Which document identities have the closest match to the set of N-grams submitted with the search?

Naturally, the sorted N-grams are not strongly consistent with the underlying things. There may be things with identities that have not yet been indexed. Sometimes, there are indices that contain the identities for things that no longer exist. *While the things and the indices may slide around, the identities usually stay intact.*

Figure 2. E-commerce—A tangled Web of identifiers.



Using Identity to Learn

Data science is based on identities, objects, and attributes. It has been used to learn surprising new things. Identities are key to its work.

Data science and observations. Data science revolves around identities. The identities have attributes. It is the manipulation of these identities and attributes and comparison with other identities that share those attributes that leads to new and deeper understanding.

Identities, objects, and attributes. When observations are made, they are stored as objects and given identities. These objects have attributes. Analyzing the objects may lead to additional attributes being added to them. Continued pattern matching on attributes over large collections of objects can lead to new attributes slapped onto the sides of the objects.

Sometimes, looking at patterns on the objects and their attributes leads to new objects showing the connections between existing objects. This will result in new identities for the new objects. So, the pattern of attributes becomes an identity in its own right, which may lead to new attributes.

Attributes on identities—rinse and repeat. It is the continuous cycle of looking at lots and lots of attributes on the objects and their identities that leads to more attributes. These new attributes are either attached to existing objects or used to generate new objects with their own independent identities.

Data science uses identities to achieve serendipitous learning.

Big Data Is Lots of Identities

Big-data systems such as MapReduce,² Apache Hadoop (<http://hadoop.apache.org>), and Apache Spark (<https://spark.apache.org>) take immutable inputs and apply functional transformations to produce immutable outputs. Because of the immutable nature of the inputs and outputs, it is easy to reason about fault tolerance when pieces of the work fail and are restarted.

Each of these big-data systems leverages the identities of data items to connect work and storage spread across many servers.

MapReduce and Apache Hadoop. These big-data systems look at the datasets they process as a bunch of key/

value pairs. Consider MapReduce and Hadoop:

- The map function of MapReduce takes a series of key/value pairs and makes a set of output key/value pairs. These output pairs may be the same as or different than the map function input.

- The reduce function is called once for each unique key and can iterate through the values associated with that key. There may be multiple values for a single key.

Queries, joins, and more with keys. Queries and joins in these big-data environments leverage the keys in the key/value pairs. These are sorted across shards with the map function. The queries and joins are applied by the reduce function handling all key/value pairs with the same key (or identity).

Because the map function can arrange an input key/value into another shaped key/value, MapReduce and Hadoop can query, sort, and join on arbitrary fields in the data. Putting the join fields into the key and sorting allows for a huge flexibility in function.

Big-data means handling lots of keys. Big-data systems require handling *lots* of keys. They can be spread around in a scalable fashion across very large clusters of servers to accomplish massive scale. The identity provided by the keys hooks it all together.

The “Internet of Identities.” IoT, or the Internet of Things, is the new trend wherein massive numbers of events from disparate devices are processed at high rates.

Internet of Things: Identifying the thing. In IoT, an extremely large number of devices that may barely qualify as computers generate massive numbers of events to be processed. Each of these devices will have an identifier in some form. As it generates events, each of these events will have a more detailed identifier that usually specifies its device of origin.

Each of these events will, in turn, have a bunch of attributes that are specific to the device. Events originating from your refrigerator will have different attributes than events originating from your car’s transmission or from a security camera at a large stadium.

Querying, joining, and connecting things. Similar to what is seen in big data, each of these IoT events has an identity and a bunch of attributes.

These events can be queried, joined, and connected based on their attributes. You can create new events by extracting attributes from a single event or from a join across multiple events.

An event is an immutable set of attributes with an identity.

The Quest for Identity

Some of today’s most challenging problems come from the quest for identity. Product matching, data science, fraud detection, homeland security, and more all struggle with figuring out when one thing is the same as another thing so identity can be assigned.

Product matching is finding identity. As discussed, providing an integrated marketplace for stuff sold by wildly disparate merchants is a big challenge. The core of this challenge is matching different SKUs from different merchants with different descriptions to find the same product identity.

This is often made easier with UPC or ISBN codes that actually do match. This leaves the product-matching system with the easier job of comparing attributes to verify identity. Product matching is not always given the boost from shared unique identifiers, and the problem becomes a task of data science.

Data science is finding identity. In data science, there are many objects, each with many attributes. Each object has a unique identity.

- Attaching new attributes: By comparing many objects and their attributes, the data-science algorithm associates new attributes with existing objects.

- Merging object identities: By examining the attributes bound to sets of objects, the data-science algorithm can realize two objects are one. That, in turn, unites their attributes.

Repeating the attribute/merge pattern causes a new understanding of identity.

Fraud detection is finding identity. Banks issuing credit cards invest heavily in fraud detection, as do retailers and other institutions that accept credit cards. Very large companies that accept credit cards have a strong incentive to detect fraud since their banks will charge them lower fees if their rate of fraud is noticeably lower. Fraud detection is big business.

Fraud detection works by looking at the transactions as objects with as-

sociated attributes. Also, credit-card holders are objects with associated attributes. Pattern-matching fraudulent activity from other credit cards to this card can give early warning. Without this matching to find new identities, e-commerce would be very challenging because of the amount of fraud that would get through.

Homeland security is finding identity. Another example of identities and matching comes from looking at patterns of travel, locations, payment types, and more. It is not unusual for an analysis of many travelers to result in similar behavior by ostensibly different people. By realizing they have the same identity, the details known about the different people can be coalesced to gain a better understanding of the risks they may pose.

Laser-sharp vision and blurring details. This coalescing of identities based upon common attributes is the basis for many of the emerging use cases in data science. One perspective is that the set of attributes defines the identity that results from the coalescing. Must the attributes be a match in all their full glory? What makes it OK to have differences? Do we want laser-sharp exactitude in the attribute matching, or is it OK to squint a little bit and blur some details to allow more matches?

Increasingly, the original data (for example, merchant feeds with product info) is kept and linked to the normalized, matched, and sanitized data. These operations are intrinsically lossy as you strive for commonality with other inputs. Considering the aligned and sanitized common view and comparing it with the individual raw feeds can offer additional insight.

Using Identity for Activities

Activities are long-running work across time and across computers, and may run across trust boundaries, departments, and companies. An activity is usually handled by having an identifier for the activity and separate identifiers for each step.

Long-running workflow runs and identifiers. Long-running workflow runs with messages across time and typically waits for external actions to complete. As external events are initiated, somehow an identifier for the event is received when it completes. To deal with an external

The judicious use of ambiguity and interchangeability lubricates distributed, long-running, scalable, and heterogeneous systems.

computer, the identifier is usually tied to outgoing and incoming messages.

Identifiers crossing trust boundaries. Sometimes an activity crosses trust boundaries. Sending messages across companies in a B2B solution opens up trust concerns—perhaps sending messages across departments or even from a Linux box to a Windows box. Each of these solutions offers challenges. The work in these cases is invariably knit together with some form of identity. That identity must have a scope in space that covers all the trusting participants and a scope in time covering the duration of the work.

It is not uncommon for one system to provide an alias for its identifiers. Messages going out and in are translated between the two identity systems.

Bank check numbers and idempotence. An example of an identifier for long-running work is the check number on the printed checks from your bank. When you make a paper check out to the electric company or the grocery store, the check has a unique identifier. On the bottom of the check are three series of numbers: the ABA (American Banking Association) routing number, account number, and check number. The ABA routing number uniquely identifies your bank. The account number identifies your account within the bank. Finally, the check number is unique within your account.

When your check is handed over to your grocery store, it is deposited in the store's bank, not yours. That bank then records the deposit along with the numbers from your check. The grocery store's bank then forwards the check to your bank, which records the debit and sends money back to the grocery store's bank.

Because of the unique identifier on the check, your bank and the grocery store's bank can implement algorithms to ensure the exactly-once processing of the debit and credit. This has been going on for many years, longer than we have had computers.

Identifiers: The glue that binds and splits. Identifiers are the glue that connects work. It's the ability to connect the work that allows us to split apart our scaling solutions and to connect previously disconnected solutions.

REST: URL-ey binding. Representa-

tional state transfer³ (REST) is an interesting and influential pattern that leverages HTTP and URLs. In the REST pattern, resources are implemented as client-server calls, which are stateless. *Stateless* means each request from the client holds enough information to process the request at the server without taking advantage of any context stored at the server. The session state is effectively held at the client.

Resources and representations. Within REST, resources are any piece of information that can be named. Typically, the name of a resource is a URL.¹ A resource is frequently used to represent groupings of related stuff that may be used to do work. The contents of a resource may be static or dynamic. What is essential is that it can be *named*.

REST resources may project one or more *representations*. Each representation is a view onto the resource that may or may not be customized for each user. The resource is itself given its own URL(s) as identities.

REST: Representational state transfer. Users wishing to work with the resource are given their own representations as identified with one or more URLs. The resource may have many users. The vast URL identity space is subdivided into representations for each user. Requests for work are accomplished with HTTP PUT commands making modifications to the representation.

Scribbling on the representation. Changing the state projected in the representation is how work is done. The combination of the representation (possibly personalized to the client) and the ability to scribble changes on the representation allows many clients to work with the resource.

URL: Mixing identity, operations, and session state. As changes to the representation occur, responses to the HTTP PUT requests are wrapped up in the URL returned. Contained in that URL is the session state describing ongoing and potentially long-running work for this client.

Identity in the URL captures the operation to be performed.

Identity in the URL captures the session state of the work!

REST: Every verb can be nounsed. REST maps a user's perspective to a set of URLs for the representation. REST also defines the mechanism for invok-

The real art of interchangeability lies in finding a way to identify the equivalent set of individuals.

ing computation and work as modifications to the representation. It's *REST* or changes to the representation that cause change.

Every verb (operation) can be nounsed (cast as data).

These nouns are described as URLs.
Operations are cast into identities represented at URLs.

Identity is RESTing on its laurels. The identity captured in the URL is a large part of why REST is so powerful. The underlying resource has an identity in the URL namespace. Each representation (assigned to a single user) has an identity in the URL namespace. Specific operations are captured, leveraging identity within the URL namespace—a powerful mechanism using the identity of the URL!

Scoping Identity

Identities must be scoped in space and time so they do not cause ambiguities. This is, on the one hand, an obvious and silly thing to say. On the other hand, it is a liberating concept.

Identifiers may have permanent unique IDs like those offered by UUIDs. These are powerful and useful. Identifiers may have a centralized or hierarchical authority that assigns their IDs, and that, by itself, offers challenges: Does this authority scale? Is it broad enough in its role to encompass the many different pieces of the solution?

The reality for most systems is that identities span the participants that see the use of that specific identifier. When merchants interact with a big e-commerce site, they will have shared identifiers for their cooperative work. Still, the merchants may not share the identifiers they use to deal with *private suppliers*. Those private suppliers may have different identifiers used to interact with the manufacturers of their products.

The scope of the identifiers is typically subject to the portion of the workflow that hosts the identifier. There are global IDs like UPC or SSN (Social Security number), but there are also local IDs like SKUs that are defined only for a single merchant.

The 'I's Have It

Identity is an extremely important part of our systems. Its real power is unleashed when combined with three

other “I” words: idempotence, immutability, and interchangeability.

Identity and idempotence. Idempotence is the property that says it’s OK to do work more than once. If it happens *at least once*, the behavior is the same as if it happens *exactly once*.⁵ In general, idempotence is a subjective concept that ignores side effects outside of the plane of abstraction provided by the service.⁶

Idempotence frequently depends on having an identity for the work. In many cases, you need to understand the identity of the operation to decide if you have done it before. There are other cases such as reading a record where the work is naturally idempotent because it leaves no effects when it’s performed. In cases where changes are made, tracking that it’s already done requires identity of some kind.

Sometimes, the identity used to provide idempotence is a consequence of some connection or session. That works until a new session arrives to retry the failed session.

Managing the requester’s identity, the target’s identity, and the identity of the work in question are some of the hardest problems in scalable systems that need idempotence.

Banks have used a simple approach to identity an idempotence with two basic tricks:

- ▶ The transaction’s identity is the preassigned check number.
- ▶ The check must typically clear in less than one year after it was written.

The second constraint limits the list of cleared checks the bank must maintain while preserving exactly-once processing.

Identity and immutability. Immutability is the property that something does not change. No matter how many times the data is read, the same result is returned. Immutability is the basis for many of today’s solutions, from low-level hardware to massively scalable solutions.⁶

Immutability is a relationship between an identity and a result that is unchanging.

Without some formalized notion of identity, you don’t have immutability.

We may see immutable identifiers for changing stuff.

A product ID may be fixed for years while its description evolves.

Identity and interchangeability. Interchangeability can be viewed as a duality with immutability. Rather than asking, “Is this thing identical?” to what we had before, we ask, “Is this thing equivalent?” to what we had before. Is it good enough?

When manufactured items are all brand new and identical, you can be happy taking any one of them from the warehouse, assuming they are not damaged. There is an identity for the product, and that identity means any one of them will do. They are interchangeable.

When reserving a room at a hotel, you accept that one king-sized non-smoking room is as good as another—even if one is next to the elevator and really noisy. The group of rooms labeled as king-sized nonsmoking is considered equivalent, and *there is an identity for any one of those rooms*. You reserve one from the pool of rooms without knowing exactly which one.

Recall that an identifier for a product description in a product catalog refers to an ambiguous version of the product description. That’s OK... any one will do, as the versions are interchangeable.

Conclusion

It used to be we focused on one application running on one computer accessing one SQL database. While we may have had application-based identifiers (for example, Social Security numbers), the underlying system was based on values in cells. *Relational algebra related values to other values.*

As systems *cleave apart* for scale, *cleave apart* to provide management or trust boundaries, or *cleave together* to integrate solutions, identifiers and identity form the glue that binds solutions. Identities also formalize the separation of disparate and distrusting solutions. Cleaving apart or cleaving together requires identities.

When we bind work together with identities, the interesting tension is, “What constitutes the identity?” What precisely is identified by a king-sized nonsmoking room? Where did we deliver the message that was guaranteed to be delivered?

New emerging systems and protocols both tighten and loosen our notions of identity, and that’s good! They make it easier to get stuff done. REST,

IoT, big data, and machine learning all revolve around notions of identity that are deliberately kept flexible and sometimes ambiguous. Notions of identity underlie our basic mechanisms of distributed systems, including interchangeability, idempotence, and immutability.

Finally, don’t be too picky about calling this identity. We see identity as names, keys, pointers, handles, IDs, numbers, identifiers, UUIDs, GUIDs, document IDs, UPCs, ASINs, employee numbers, file names, Social Security numbers, and much more.

Truly, identity by any other name *does smell as sweet ...* □

Related articles on queue.acm.org

Pervasive, Dynamic Authentication of Physical Items

Meng-Day Yu and Srinivas Devadas
<https://queue.acm.org/detail.cfm?id=3047967>

How Do I Model State? Let Me Count the Ways

Ian Foster et al.
<https://queue.acm.org/detail.cfm?id=1516638>

How to De-identify Your Data

Olivia Angiuli, Joe Blitzstein, and Jim Waldo
<https://queue.acm.org/detail.cfm?id=2838930>

References

1. Berners-Lee, T., Masinter, L. and McCahill, M. Universal Resource Locator. Technical Report, Internet Engineering Task Force, Draft RFC, 1994; <https://dl.acm.org/citation.cfm?id=RFC1738>.
2. Dean, J. and Ghemawat, S. MapReduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation* 6, 10 (2004); <https://dl.acm.org/citation.cfm?id=1251264>.
3. Fielding, R. Architectural styles and the design of network-based software. Ph.D. dissertation. University of California, Irvine, 2000.
4. Helland, P. The power of babble. *Commun. ACM* 59, 11 (Nov. 2016), 40–43; <https://dl.acm.org/citation.cfm?id=2980932>.
5. Helland, P. Idempotence is not a medical condition. *acmqueue* 10, 4 (2012); <https://dl.acm.org/citation.cfm?id=2187821>.
6. Helland, P. Immutability changes everything. *acmqueue* 13, 9 (2016); <https://queue.acm.org/detail.cfm?id=2884038>. (First printed in the Biennial Seventh Conference on Innovative Database Research (January 2015).
7. Helland, P. Data on the outside versus data on the inside. In *Proceedings of the Conference on Innovative Database Research*; <http://cidrdb.org/cidr2005/papers/P12.pdf>.
8. Helland, P. Side effects, front and center! *Commun. ACM* 60, 7 (July 2017), 36–39; <https://dl.acm.org/citation.cfm?id=3080010>.

Pat Helland has been implementing transaction systems, databases, application platforms, distributed systems, fault-tolerant systems, and messaging systems since 1978. He currently works at Salesforce.



Critical but oft-neglected service metrics that every SRE and product owner should care about.

BY BENJAMIN TREYNOR SLOSS, SHYLAJA NUKALA, AND VIVEK RAU

Metrics That Matter

SITE RELIABILITY ENGINEERING, or SRE, is a software-engineering specialization that focuses on the reliability and maintainability of large systems. In its experience in the field, Google has found some critical but oft-neglected metrics that are important for running reliable services.

This article, based on Ben Treynor's talk at the Google Cloud Next 2017 conference,⁷ addresses those metrics, specifically for product development and SRE teams, managers of such teams, and anyone else who cares about the reliability of Web products or infrastructure. To further explain its approach to product reliability, Google has published *Site Reliability Engineering: How Google Runs Production Systems*¹ (hereafter referred to as the SRE book) and

*The Site Reliability Workbook: Practical Ways to Implement SRE*² (hereafter referred to as the SRE workbook).

One of the most important choices in offering a service is which service metrics to measure, and how to evaluate them. The difference between great, good, and poor metric and metric threshold choices is frequently the difference between a service that will surprise and delight its users with how well it works, one that will be acceptable for most users, and one that will actively drive away users—regardless of what the service actually offers.

For example, it is not uncommon to measure the QPS (queries per second) received at a Web or API server, and to assess that this metric indicates good service health if the graph of the metric over time has a smooth sinusoidal diurnal curve with no unexpected spikes or troughs, and the peaks of the curve are rising over time, indicating user growth. Yet this is a poor metric choice—at best it will provide the operator with a lagging indicator of large-scale problems. It misses a host of real, common problems, including partial unreachability, error rates in the 0.1%–3% range, high latency, and intervals of bad results.

These problems lead to unhappy users and service abandonment—yet throughout it all, the QPS Received graph continues to show its happy sinusoidal curves and to provide a soothing sense that all is well. The best that can be said about the QPS Received metric is that it's relatively simple to implement—and even that is a problem, because it is often implemented early and thus takes the place of more sophisticated and useful metrics that would provide an operator with more accurate and useful data about the service.

What follows are the types of metrics the Google SRE team has adopted for Google services. These metrics are not particularly easy to implement, and they may require changes to a service to instrument properly. It has been our consistent experience at Google, however, that every service team that



implements these metrics is happy afterward that it made the effort to do so. The metrics investment is small compared with the overall effort to build and launch the service in the first place, and the prompt payback in user satisfaction and usage growth is outsized relative to the effort required. We believe you will find this is true for your service, too.

Lesson 1. Measure the Actual User Experience

The SRE book emphasizes that speed matters to users, as demonstrated by Google's research on shifts in behavior when users are exposed to delayed responses from a Web service.³ When services get too slow, users start to disengage, and when they get even slower, users leave. "Speed matters" is a good axiom for SREs to apply when thinking about what makes a service attractive to users.

A good follow-up question is, "Speed for whom?" Engineers often think about measuring speed on the server side, because it is relatively easy to instrument servers to export the required metrics, and standard monitoring tools are designed to capture such metrics from servers in dashboards and highlight anomalies with alerts. What this standard setup is measuring is the interval between the point in time when a user request enters a datacenter and the point in time when a response to that request leaves the datacenter. In other words, the metric being captured is server-side latency. Measuring server-side latency is not sufficient, though it is better than not measuring latency at all. Measuring and reporting on server-side latency can be a useful stopgap while solving the harder problem of measuring client-side latency.

The problem is that users have no interest in this server-side metric. Users care about how fast or slow the application is when responding to their actions, and, unfortunately, this can have very little correlation with server-side latency. Perhaps these users have a cheap phone, on a slow 2G network, in a country far away from your servers; if your product doesn't work for them, all your hard work building great features will be wasted, because users will be unhappy and will use a different product. The problem will be compounded

Though difficult, client-side metrics are essential and achievable.

if you are measuring only server-side latency, because you will be completely unaware the product is slow for users. Even if you get anecdotal reports of slowness and try to follow up on them, you will have no way of determining which subset of users is experiencing slowness, and when.

To measure the actual user experience, you have to measure and record client-side latency. It can be hard work to instrument the client code to capture this latency metric and then to ship client-side metrics back to the datacenter for analysis. The work may be further complicated by the need to handle broken network connections by storing the data and uploading it later.

Though difficult, client-side metrics are essential and achievable.

For a browser application, you can write additional JavaScript that gathers these statistics for users on different platforms, in different countries, and so on, and send these statistics back to the server. For a thick client, the path is more obvious, but it's still important to measure the time from the moment the user interacts with the client until the response is delivered. Either way, instrumenting the user experience takes a relatively small fraction of the effort previously expended to write the entire application, and the payback for this incremental effort is high.

To take an example from Google's own history, when Gmail was launched, most users accessed it through a Web browser (not a mobile client), and Google's Web client code had no instrumentation to capture client-side latency. So, we relied on server-side latency data, and the response time seemed quite acceptable. When Google finally launched an instrumented JavaScript client, at first we did not believe the data it was sending back—it seemed impossible the user experience was that bad. We went through the denial stage for a while, and then anger, and eventually got to bargaining.⁴ We made some major changes to how the Gmail server and its client worked to improve our client-side latency, and the reward was a visible inflection point in Gmail's growth once the user experience improved. The long-term trends in our monitoring dashboards showed users responding to the improved product experience. For around 3% of the effort

of writing and running Gmail, there was a major increase in its adoption and user happiness.

Many techniques are available to application developers for improving client-side response times, and not all of them require large engineering investments. Google's PageSpeed project was created to share with the world the company's insights into client-side response optimization, accompanied by tools that help engineers apply these insights to their own products and Web pages.⁵ One of the obvious rules is to reduce server response time as much as possible. PageSpeed analysis tools also recommend various well-known techniques for client-side optimization, including compression of static content, using a preprocessor to "minify" code (HTML, CSS, and JavaScript) by removing unnecessary and redundant text, setting cache-control headers correctly, compressing or inlining images, and more.

To recap, measure the actual user experience by measuring how long a user must wait for a response after performing an action on your product. Do this, even though it is often not easy. Experience says it will be well worth the effort.

Lesson 2. Measure Speed at the 95th and 99th Percentiles

While "Speed matters" is a good axiom when thinking about user (un)happiness, that still leaves an open question about how best to quantify the speed of a service. In other words, even if you understand and accept the value of the latency metric (time to respond to user requests) should be low enough to keep users happy, do you know precisely what metric that is? Should you measure average latency, median latency, or nth-percentile latency?

In the early days of Google's SRE organization, when we managed relatively few products other than Search and Ads, SLOs (service-level objectives) were set for speed based on median latency. (An SLO is a target value for a given metric, used to communicate the desired level of performance for a service. When the target is achieved, that aspect of the service is considered to be performing adequately. In the context of SLOs, the

How to Define Percentile-Based SLOs

There is a technique to phrasing SLO definitions optimally—a linguistic point illustrated here with an amusing puzzle. Consider these two alternative SLO definitions for a given Web service, using slightly different language in each definition:

1. The 99th-percentile latency for user requests, averaged over a trailing five-minute time window, will be less than 800 milliseconds.
2. Some 99% of user requests, averaged over a trailing five-minute time window, will complete in less than 800 milliseconds.

Assume the SLO will be measured every 10 seconds in either case, and an alert will be fired if N consecutive measurements are out of range. Before reading further, think about which SLO definition is better, and why.

The answer is that from a user-happiness perspective, the two SLOs are practically equivalent; and yet, from a computational perspective, alternative number 2 is distinctly superior.

To appreciate this, consider a hypothetical Web service receiving 10,000 user requests per second, on average, under peak load conditions. With SLO definition 1, the measurement algorithm actually has to compute a percentile value every 10 seconds. A naive approach to this computation is as follows:

- ▶ Store the response times for $10,000 \times 300 = 3$ million queries in memory to capture five minutes' worth of data (this will use >11MB of memory to store 3 million 32-bit integers, each representing the response time for one query in milliseconds).
 - ▶ Sort these 3 million integer values.
 - ▶ Read the 99th-percentile value (that is, the 30,000th latency value in the sorted list, counting from the maximum downward).

More efficient algorithms are definitely available, such as using 16-bit short integers for latency values and using two heaps instead of sorting a linear list every 10 seconds, but even these improved approaches involve significant overhead.

In contrast, SLO definition 2 requires storing only two integers in memory: the count of user requests with completion times greater than 800 milliseconds, and the total count of user requests. Determining SLO compliance is then a simple division operation, and you don't have to remember latency values at all.

Be sure to define your long-tail latency SLOs using format 2.

metric being evaluated is called an SLI, or service-level indicator.)

Over the years, particularly as the use of Search expanded to other continents, we learned that users could be unhappy even when we were meeting and beating our SLO targets. We then conducted research to determine the impact of slight degradations in response time on user behavior, and found that users would conduct significantly fewer searches when encountering incremental delays as small as 200 milliseconds.³ Based on these and other findings, we have learned to measure "long-tail" latency—that is, latency must be measured at the 95th and 99th percentiles to capture the user experience accurately. After all, it doesn't matter if a product is serving the correct result 99.999% of the time if 5% of users are unhappy with how long it takes to get that correct result.

Once upon a time, Google measured only raw availability. In fact, most SLOs even today are framed around availability: How many requests return a good result versus how many return an

error. Availability was computed the following way:

$$\% \text{ Availability} = 1 - \% \text{ error responses}$$

Suppose you have a user service that normally responds in half a second, which sounds good enough for a user on a smartphone, given typical wireless network delays. Now suppose one request in 30 has an internal problem causing a delay that leads to the mobile client app retrying the request after 10 seconds. Now further suppose the retry almost always succeeds. The availability metrics (as computed here) will say "100% availability." Users will say "97% available"—because if they are accustomed to receiving a response in 500 milliseconds, after three to five seconds they will hit retry or switch apps. It doesn't matter if the user documentation says, "The application may take up to 10 seconds to respond;" once the user base is trained to get an answer in 500 milliseconds most of the time, that is what they will expect, and they will be

have like a 10-second response delay is an outage. Meanwhile, the SREs will (incorrectly) be happy, at least for the time being, because their measurements say the service is 100% available. This disconnect can be avoided by correcting the availability computation as follows:

$$\% \text{ Availability} = 1 - \% (\text{error responses} + \text{slow responses})$$

Therefore, when an SLO is defined for long-tail latency, you must choose a target response time that does not render the service effectively unavailable. The 99th-percentile latency should be such that users experiencing that latency do not find it completely unacceptable relative to their expectations. Note their expectations were probably set by the median latency. You really do need to know what

your users consider minimally acceptable. A good practice is to conduct experiments that measure how many users are actually lost as latency is artificially increased. These experiments should be conducted infrequently, using a tiny fraction of randomly sampled users to minimize the risk to your product's brand and reputation.

A good practical rule of thumb learned from these experiments at Google is that the 99th-percentile latency should be no more than three to five times the median latency. This means that if a hypothetical service with median latency of 400 milliseconds starts exhibiting more than two seconds response time for the slowest 1% of requests, this is undesirable. We tune our production systems such that if this undesired behavior continues for some predefined period, an alert will fire or some automated corrective action will be taken (such as shifting traffic around or provisioning more servers). We find the 50th-, 95th-, and 99th-percentile latency measures for a service are each individually valuable, and we will ideally set SLOs around each of them.

Our recommendations for latency metrics can be applied equally well to other kinds of SLIs, some of them applicable to systems that are not Web services. As discussed in the SRE book, storage systems also care about *durability* (whether data is available when needed), and data-processing pipelines care about throughput and *freshness* (how long it takes for data to progress from ingestion to completion).^a

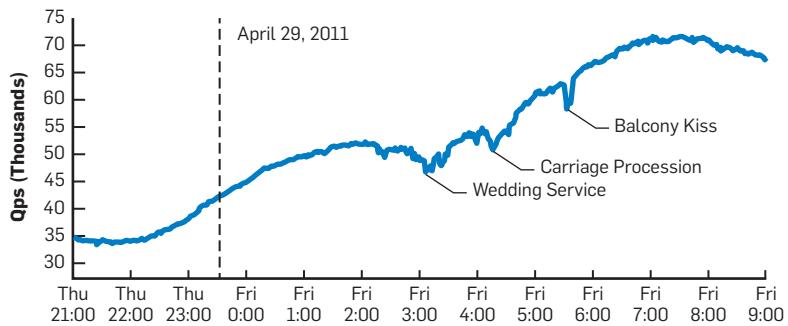
Lesson 3. Measure Future Load

Demand forecasting, or quantifying the future load on a service, is different from typical SLO measurement because it's not a metric you monitor, nor a cause for generating alerts. Demand forecasting makes a service reliable by providing the information needed to provision the service such that it can handle its future load while continuing to meet its SLOs. The more effort you put into generating good demand forecasts, the less you will need to scramble at the last minute to add more compute resources to the service

^a For more advice on how to create SLOs for a service, read "Implementing SLOs," in the SRE workbook.

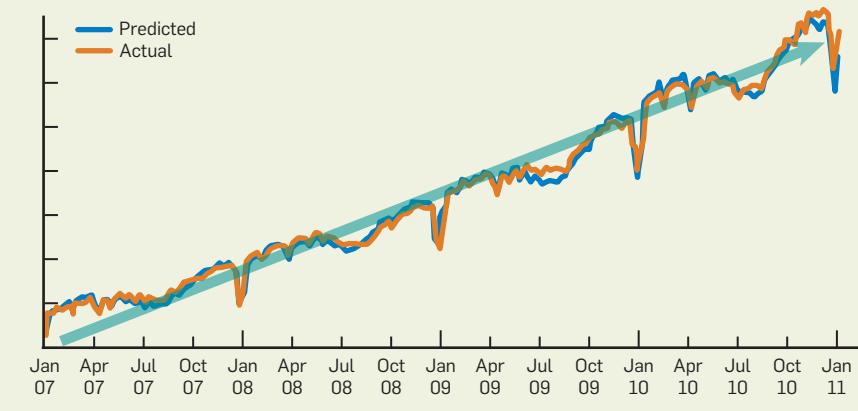
The Royal Wedding as Seen by Google Search Monitoring

Royal wedding searches.



Daily traffic fluctuations are far less important for capacity planning than monthly or yearly increases, but they provide an amusing illustration of the impact of external world events on the load presented to a Web service. The figure here was generated by the system that monitors load on Google's Search product and represents the number of search QPS on April 29, 2011, during the wedding of Prince William and Kate Middleton. The time values on the y-axis are in the Pacific time zone (eight hours behind U.K. time), and the traffic pattern neatly captures key events during the ceremony. It is evident from charts like this one that when something really interesting happens in the world, people briefly stop searching the Web, and when that event is over, they promptly resume searching.

Average weekly Google search traffic.



because it's melting down in the face of an unforeseen increase in traffic.

Load on a service is measured using different combinations of metrics depending on the type of service being discussed, but a common denominator unit for many services is QPS. Layered on top of QPS might be other service-dependent metrics such as storage size (gigabytes or terabytes), memory usage, network bandwidth, or I/O bandwidth (gigabits per second).

It's useful to break demand growth down into *organic* and *inorganic*. Organic growth is what you can forecast by extrapolating historical trends in traffic, and the forecasting problem can often be addressed using statistical tools. Inorganic growth is what you forecast for one-time events such as product launches, changes in service performance, or anticipated changes in user behavior, among other factors, and this growth cannot be extrapolated from historical data. Prediction of inorganic growth is less amenable to statistical tools and often relies on rules of thumb and estimates derived from similar events in the past. In the time leading up to a service launch, when there is not enough historical data available to make an organic growth forecast, teams estimate demand using techniques applicable to inorganic growth.

Forecasting organic growth. For mature products that have been in operation for a few years, you can forecast organic growth using statistical methods. Note that linear regression is not a useful tool in most cases, because it does not capture seasonal traffic fluctuations; it also does not work if growth is not linear. Many Web services see significant drops in traffic (the "summer slump") because of the midyear vacation season, and, conversely, see big spikes in traffic during the year-end shopping season, followed by a major "holiday dip" in the last week of the year, followed in turn by a "back-to-work bounce" at the start of the new year (see the accompanying figure). At Google, we even account for predictable changes with a cycle time of several years, caused by events such as the FIFA World Cup.

Google uses a variety of forecasting models that attempt to capture seasonality on a monthly or annual

Google Analytics Lesson Learned

An interesting case study of inorganic growth that was not generated by any engineering change and was not small involves the initial launch of Google Analytics, a service for gathering and analyzing traffic to any website. Google had acquired Urchin Software Corporation for its Web-analytics product that provided traffic collection and analytics dashboards to paying customers. The inorganic traffic growth event occurred when the product was made available for free under the Google brand, permitting any website owner to sign up for it at no charge. Google correctly anticipated a flood of new users, based on prior experience launching the Keyhole (later called Google Earth) subscription-based product for free. Therefore, we carefully load tested and provisioned the product for the expected increase in traffic.

Our prediction for core product usage then performed reasonably well, but we had forgotten to account for traffic to the signup page! The page where new users signed up was backed by a single-threaded SQL database with limited transaction capacity, placing a strict and previously unknown limit on the number of signups per second, resulting in a stream of public complaints from users about site slowness and unavailability. We learned this lesson well, and our product launch checklist afterward contained the question, "*Do new users have to sign up for your service, and if so, have you estimated and tested the load on your signup page?*"

time scale. There is uncertainty in forecasts, and they imply a confidence level, so rather than forecasting a line, we are forecasting a cone. Any given statistical model has its strengths and weaknesses, so many Google products use outputs generated from a large ensemble of models,⁶ which include variants on many well-known approaches, such as the Bass Diffusion Model; Theta Model; logistic models; Bayesian Structural Time Series; STL (seasonal and trend decomposition using Loess); Holt-Winters and other exponential smoothing models; seasonal and other ARIMA (autoregressive integrated moving average)-based models; year-over-year growth models; custom models; and more.

Having generated independent estimates from each model in the ensemble, we then compute their mean after applying a configurable "trimming" parameter to eliminate outlier estimates, and this adjusted mean is used as the final prediction. Depending on the scale and global reach of a service and its different levels of adoption in different parts of the world, it might be more accurate to generate continent-level or country-level forecasts and aggregate them instead of attempting to forecast at the global level.

It is important to compare forecasts regularly with actual traffic in order to tune the model parameters over time and improve the accuracy of the models. Experience shows that

the trimmed mean of the ensemble of models delivers superior accuracy compared with any individual model.

Forecasting inorganic growth. Inorganic growth is generated by one-time events that have no periodicity, such as launches of new products, new features, or marketing promotions, or changes in user behavior that are triggered by some extraneous factor for which the timing is predictable but the resulting peak traffic volume has a high degree of uncertainty (like the FIFA World Cup or the Royal Wedding), among others. Inorganic growth involves an abrupt change in traffic and is intrinsically unpredictable because it is triggered by an event that hasn't happened before or otherwise cannot be directly extrapolated from the past. When the product owners and SREs have advance notice of such growth, such as when planning for a new feature launch, they need to apply intuition and rules of thumb to estimating post-launch traffic, and understand their predictions will have a higher level of uncertainty.

General rules for forecasting inorganic growth for product/feature launches include the following:

- ▶ Examine historical traffic changes from past launches of similar or analogous features.
- ▶ For country- or market-specific launches, consider past user behavior in that market.
- ▶ Consider the level of publicity and promotion around the launch.

► Add a margin of uncertainty to the forecast where possible, by provisioning three to five times the resources implied by the forecast.

► While traffic from brand-new products is harder to predict, it is also usually small, so you can overprovision for this traffic without incurring too much cost.

Lesson 4. Measure Service Efficiency

SRE teams should regularly measure the efficiency of each service they run, using load tests and benchmarking programs to determine how many user requests per second can be handled with acceptable response times, given a certain quantity of computing resources (CPU, memory, disk I/O, and network bandwidth). While performance testing may seem an obvious best practice, in real life teams frequently forget about service efficiency. They may benchmark a service once a year, or just before a major release, and then assume unconsciously that the service's performance remains constant between benchmarks. In reality, even minor changes to the code, or to user behavior, can affect the amount of resources required to serve a given volume of traffic.

A common way of finding out that a service has become less efficient is through a product outage. The SRE team may think they have enough capacity to serve peak traffic even with two datacenters' worth of resources turned down for maintenance or emergency repairs, but when the rare event occurs where both datacenters are actually down during peak traffic hours, the performance of the service radically degrades and causes a partial outage or becomes so slow as to make the service unusable. In the worst case, this can turn into a "cascading failure" where all serving clusters collapse like a row of dominoes, inducing a global product outage.

Ironically, this type of massive failure is triggered by the system's attempt to recover from smaller failures. One cluster of servers happens to get a higher load for reasons of geography and/or user behavior, and this load is large enough to cause all the servers to crash. The traffic load-balancing system observes these servers going off-

line and performs a failover operation, diverting all the traffic formerly going to the crashed cluster and sending it to nearby clusters instead. As a result, each of these nearby servers now gets even more overloaded and crashes as well, resulting in more traffic being sent to even fewer live servers. The cycle repeats until every single server is dead and the service is globally unavailable.

Services can avoid cascading failures using the *drop overload* technique. Here the server code is designed to detect when it is overloaded and randomly drop some incoming requests under those circumstances, rather than attempting to handle all requests and eventually melting down. This results in a degraded customer experience for users whose requests are dropped, but that can be mitigated to a large extent by having the client retry the request; in any case, slower responses or outright error responses to a fraction of users are a lot better than a global service failure.

It would be better, of course, to avoid this situation altogether, and the only way to do that is to regularly measure service efficiency to confirm the SRE team's assumptions about how much serving capacity is available. For a service that ships out releases daily or more frequently, daily benchmarking is not an extreme practice—benchmarking can be built into the automated release testing procedure. When newly introduced performance regressions are detected early, the team can provision more resources in the short term and then get the performance bugs fixed in the long term to bring resource costs back in line.

If you run your service on a cloud platform, some cloud providers have an *autoscaling* service that will automatically provision more resources when your service load increases. This setup may be better than running products on premises or in a datacenter with fixed hardware resources, but it still does not get you off the hook for regular benchmarking. Even though the risk of a complete outage is lower, you may find out too late that your monthly cloud bill has increased dramatically just because someone modified the encoding scheme used for compressing data, or made some other seemingly innocuous code change. For

these reasons, it is a best practice to measure service efficiency regularly.^b

Conclusion

The metrics discussed in this article should be useful to those who run a service and care about reliability. If you measure these metrics, set the right targets, and go through the work to measure the metrics accurately, not as an approximation, you should find that your service runs better; you experience fewer outages; and you see a lot more user adoption. Most of us like those three properties. □

- b For additional details, see "Managing Load," in the SRE workbook. Chapter 11 contains two case studies of managing overload.

Related articles on queue.acm.org

A Purpose-built Global Network: Google's Move to SDN

A discussion with Amin Vahdat, David Clark, and Jennifer Rexford
<https://queue.acm.org/detail.cfm?id=2856460>

From Here to There, the SOA Way

Terry Coatta
<https://queue.acm.org/detail.cfm?id=1388788>

Voyage in the Agile Memeplex

Philippe Kruchten
<https://queue.acm.org/detail.cfm?id=1281893>

References

1. Beyer, B., Jones, C., Petoff, J. and Murphy, N.R. *Site Reliability Engineering: How Google Runs Production Systems*. O'Reilly Media, 2016.
2. Beyer, B., Murphy, N.R. and Rensin, D.K., Kawahara, K., Thorne, S. *The Site Reliability Workbook: Practical Ways to Implement SRE*. O'Reilly Media, 2018.
3. Brutlag, J. Speed matters. Google AI Blog, 2009; <https://research.googleblog.com/2009/06/speed-matters.html>.
4. Kübler-Ross, E. Kübler-Ross Model; https://en.wikipedia.org/wiki/K%C3%BCbler-Ross_model.
5. PageSpeed. Analyze and optimize your website with PageSpeed tools; <https://developers.google.com/speed/>
6. Tassone, E. and Rohani, F. Our quest for robust time series forecasting at scale. The Unofficial Google Data Science Blog;
7. <http://www.unofficialgoogledatascience.com/2017/04/our-quest-for-robust-time-series.html>.
8. Treynor, B. Metrics that matter (Google Cloud Next), 2017; <https://youtu.be/iF9NoqYBb4U>.

Benjamin Treynor Sloss started programming at age 6 and joined Oracle as a software engineer at 17. He has also worked at Versant, EpiPhany, SEVEN, and (currently) Google. His team of approximately 4,700 is responsible for site reliability engineering, networking, and datacenters worldwide.

Shylaja Nukala is a technical writing lead for Google Site Reliability Engineering. She leads the documentation, information management, and select-training efforts for SRE, Cloud, and Google engineers.

Vivek Rau is a site reliability engineer at Google, working on customer reliability engineering (CRE). The CRE team teaches customers core SRE principles, enabling them to build and operate highly reliable products on the Google Cloud Platform.

Copyright held by authors/owners.

Scaling resources within multiple administrative domains.

BY NITESH MOR

Research for Practice: Edge Computing



CLOUD COMPUTING, A term that elicited significant hesitation and criticism at one time, is now the de facto standard for running always-on services and batch-computation jobs alike. In more recent years, the cloud has become a significant enabler for the IoT (Internet of

Things). Network-connected IoT devices—in homes, offices, factories, public infrastructure, and just about everywhere else—are significant sources of data that must be handled and acted upon. The cloud has emerged as an obvious support platform because of its cheap data storage and processing capabilities, but can this trend of relying exclusively on the cloud infrastructure continue indefinitely?

For the applications of tomorrow, computing is moving out of the silos of far-away datacenters and into everyday lives. This trend has been called edge computing (<https://invent.ge/2BIhzQR>), fog computing (<https://bit.ly/2eYXUxj>), cloudlets (<http://elijah.cs.cmu.edu/>), as well as other designations. In this article, edge computing serves as an umbrella term for this trend. While cloud computing

infrastructures proliferated because of flexible pay-as-you-go economics and the ability to outsource resource management, edge computing is a growing trend to satisfy the needs of richer applications by enabling lower latency, higher bandwidth, and improved reliability. Plus, both privacy concerns and legislation that require data to be confined to a specific physical infrastructure are also driving factors for edge computing.

It is important to note that edge computing is not merely caching, filtering, and preprocessing of information using onboard resources at the source/sink of data—the scope of edge computing is much broader. Edge computing also includes the use of networked resources closer to the sources/sinks of data. In an ideal world, resources at the edge, in the

cloud, and everywhere in between form a continuum. Thus, for power users such as factory floors, city infrastructures, corporations, small businesses, and even some individuals, edge computing means making appropriate use of on-premises resources together with their current reliance on the cloud. In addition to existing cloud providers, a large number of smaller but more optimally located service providers, that handle the overflow demand from power users as well as support novice users, are likely to flourish.

Creating edge computing infrastructures and applications encompasses quite a breadth of systems research. Let's take a look at the academic view of edge computing and a sample of existing research that will be relevant in the coming years.

A Vision For Edge Computing: Opportunities And Challenges

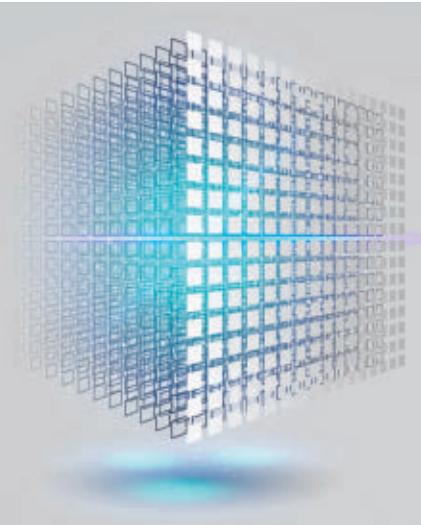
Let's start with an excellent paper that introduced the term *fog computing* and highlights why practitioners should care about it:

Fog Computing and Its Role in the Internet of Things

F. Bonomi, R. Milti, J. Zhu, and S. Addepalli
In *Proceedings of the First Edition of the ACM Workshop on Mobile Cloud Computing*, (2012) 3–16; <https://dl.acm.org/citation.cfm?id=2342513>

Although short, this paper provides a clear characterization of fog computing and a concise list of the opportunities it provides. It then goes deeper into the discussion of richer applications and services enabled by fog computing, such as connected vehicles, smart grid, and wireless sensor/actuator networks. The key takeaway is that the stricter performance/QoS requirements of these rich applications and services need better architectures for compute, storage, and networking; and appropriate orchestration and management of resources.

While this paper is specifically about the Internet of Things and fog computing, the same ideas apply to edge computing in the broader sense. Unsurprisingly, the opportunities of edge computing also come with a number of challenges. An in-depth case study of some of these challenges and possible



workarounds is illustrated in the FarmBeats project that was discussed in the RfP “Toward a Network of Connected Things” featured in the July 2018 issue of *Communications*, p. 52–54.

A World Full of Information: Why Naming Matters

One of the hurdles in using resources at the edge is the complexity they bring with them. What can be done to ease the management complexity? Are existing architectures an attempt to find workarounds for some more fundamental problems?

Information-centric networks (ICNs) postulate that most applications care only about information, but the current Internet architecture involves shoehorning these applications into a message-oriented, host-to-host network. While a number of ICNs have been proposed in the past, a recent notable paper addresses named data networking (NDN).

Named Data Networking

L. Zhang et al.
ACM SIGCOMM Computer Communication Rev. 44, 3 (2014), 66–63; <https://dl.acm.org/citation.cfm?id=2656887>

NDN, like many other ICNs, considers named information as a first-class citizen. Information is named with human-readable identifiers in a hierarchical manner, and the information can be directly accessed by its name instead of through a host-based URL scheme.

As for the architecture of the routing network itself, NDN has two types of packets: interest and data. Both

types are marked with the name of the content. A user interested in specific content creates an interest packet and sends it into the network. The NDN routing protocol is based on a name-prefix strategy, which, in some ways, is similar to prefix aggregation in IP routing. An NDN router, however, differs from IP routers in two important ways: it maintains a temporary cache of data that it has seen so far, so that any new interests from downstream nodes can be responded to directly without going to an upstream router; and only one request is sent to the upstream router for multiple interests to the same name by a number of downstream nodes. Multiple paths for the same content are also supported.

The security in NDN is also data-centric. The producer of the data cryptographically signs each data packet, and a consumer can reason about data integrity and provenance from such signatures. In addition, encryption of data packets can be used to control access to information.

Using human-readable names allows for the creation of predictable names for content, which is useful for a certain class of applications. The paper also describes how applications would look with NDN by using a number of examples such as video streaming, real-time conferencing, building automation systems, and vehicular networking.

NDN is not the first ICN, and it isn't the last. Earlier ICNs were based on flat cryptographic identifiers for addresses, compared to NDN's hierarchical human-readable names. A more detailed overview of ICNs, their challenges, commonalities, and differences can be found in a 2011 survey paper by Ghodsi et al. (<https://dl.acm.org/citation.cfm?id=2070563>).

To provide a little historical context, NDN was one of several future Internet architectures (FIAs) funded by the National Science Foundation. It is instructive to look at a few other projects, such as XIA (<https://dl.acm.org/citation.cfm?id=2070564>) and MobilityFirst (<https://dl.acm.org/citation.cfm?id=2089017>), which share the goals of cleaner architectures for the future Internet.

The key lesson for practitioners is that choosing the right level of abstrac-

tions is important for ensuring appropriate separation of concerns between applications and infrastructure.

Securing Execution

While information management is important, let's not forget about computation. Whereas cryptographic tools can help secure data, it is equally important to secure the computation itself. As containers have risen in popularity as a software distribution and lightweight execution environment, it is important to understand their security implications—not only for isolation among users, but also for protection from platform and system administrators.

SCONE: Secure Linux Containers with Intel SGX

S. Arnaudov et al.

Operating Systems Design and Implementation
16 (Nov. 2016) 689–703; <https://www.usenix.org/system/files/conference/osdi16/osdi16-arnaudov.pdf>

SCONE implements secure application execution inside Docker (www.docker.com) using Intel SGX, (<https://software.intel.com/en-us/sgx>), assuming one trusts Intel SGX and a relatively small TCB (trusted computing base) of SCONE. Note that system calls cannot be executed inside an SGX enclave itself and require expensive enclave exits. The ingenuity of SCONE is in making existing applications work with acceptable performance without source code modification, which is important for real-world adoption.

While this paper is quite detailed and instructive, here is a very brief summary of how SCONE works: An application is compiled against the SCONE library, which provides a C standard library interface. The SCONE library provides “shielding” of system calls by transparently encrypting/decrypting application data. To reduce the performance degradation, SCONE also provides a user-level threading implementation to maximize the time threads spend inside the enclave. Further, a kernel module makes it possible to use asynchronous system calls and achieve better performance; two lock-free queues handle system call requests and responses, which minimizes enclave exits.

Integration with Docker allows for easy distribution of packaged software.



The key lesson for practitioners is that choosing the right level of abstractions is important for ensuring appropriate separation of concerns between applications and infrastructure.



The target software is included in a Docker image, which may also contain secret information for encryption/decryption. Thus, Docker integration requires protecting the integrity, authenticity, and confidentiality of the Docker image itself, which is achieved with a small client that is capable of verifying the security of the image based on a startup configuration file. Finally, the authors show SCONE can achieve at least 60% of the native throughput for popular existing software such as Apache, Redis, and memcached.

While technologies such as Intel SGX do not magically make applications immune to software flaws (as has been demonstrated by Spectre (<https://bit.ly/2MzW0Xb>) and Foreshadow (<https://foreshadowattack.eu>), hardware-based security is a step in the right direction. Computing resources on the edge may not have physical protections as effective as those in cloud datacenters, and consequently, an adversary with physical possession of the device is a more significant threat in edge computing.

For practitioners, SCONE demonstrates how to build a practical secure computation platform. More importantly, SCONE is not limited to edge computing; it can also be deployed in existing cloud infrastructures and elsewhere.

A Utility Provider Model of Computing

Commercial offerings from existing service providers, such as Amazon’s AWS IoT GreenGrass and AWS Snowball Edge, enable edge computing with on-premises devices and interfaces that are similar to current cloud offerings. While using familiar interfaces has some benefits, it is time to move away from a “trust based on reputation” model.

Is there a utility-provider model that provides verifiable security without necessarily trusting the underlying infrastructure or the provider itself? Verifiable security not only makes the world a more secure place, it also lowers the barrier to entry for new service providers that can compete on the merits of their service quality alone.

The following paper envisions a cooperative utility model where users pay a fee in exchange for access to persistent

storage. Note that while many aspects of the vision may seem like a trivial task with the cloud computing resources of today, this paper predates the cloud by almost a decade.

OceanStore: An Architecture for Global-Scale Persistent Storage

J. Kubiatowicz et al.

ACM SIGOPS Operating Systems Review 34, 5 (2000), 190–201; <https://dl.acm.org/citation.cfm?id=357007>

OceanStore assumes a fundamentally untrusted infrastructure composed of geographically distributed servers and provides storage as a service. Data is named by globally unique identifiers (GUIDs) and is portrayed as nomadic (that is, it can flow freely and can be cached anywhere, anytime). The underlying network is essentially a structured P2P (peer-to-peer) network that routes data based on the GUIDs. The routing is performed using a locality-aware, distributed routing algorithm.

Updates to the objects are cryptographically signed and are associated with predicates that are evaluated by a replica. Based on such evaluation, an update can be either committed or aborted. Further, these updates are serialized by the infrastructure using a primary tier of replicas running a Byzantine agreement protocol, thus removing the trust in any single physical server or provider. A larger number of secondary replicas are used to enhance durability. In addition, data is replicated widely for archival storage by using erasure codes.

While OceanStore has a custom API to the system, it provides “facades” that could offer familiar interfaces—such as a filesystem—to legacy applications. This is a vision paper with enough details to convince readers that such a system can actually be built.

In fact, OceanStore had a follow-up prototype implementation named Pond (<https://bit.ly/2SAIjje>). In a way, OceanStore can be considered a two-part system: An information-centric network underneath and a storage layer on top that provides update semantics on objects. Combined with the secure execution of Intel SGX-like solutions, it should be possible, in theory, to run end-to-end secure applications.

Although OceanStore appeared be-

While using familiar interfaces has some benefits, it is time to move away from a “trust based on reputation” model.

fore the cloud was a thing, the idea of a utility model of computing is more important than ever. For practitioners of today, OceanStore demonstrates that it is possible to create a utility-provider model of computing even with a widely distributed infrastructure controlled by a number of administrative entities.

Final Thoughts

Because edge computing is a rapidly evolving field with a large number of potential applications, it should be on every practitioner’s radar. While a number of existing applications can benefit immediately from edge computing resources, a whole new set of applications will emerge as a result of having access to such infrastructures. The emergence of edge computing does not mean cloud computing will vanish or become obsolete, as there will always be applications that are better suited to being run in the cloud.

This article merely scratches the surface of a vast collection of knowledge. A key lesson, however, is that creating familiar gateways and providing API uniformity are merely facades; infrastructures and services are needed to address the core challenges of edge computing in a more fundamental way.

Tackling management complexity and heterogeneity will probably be the biggest hurdle in the future of edge computing. The other big challenge for edge computing will be data management. As data becomes more valuable than ever, security and privacy concerns will play an important role in how edge computing architectures and applications evolve. In theory, edge computing makes it possible to restrict data to specific domains of trust for better information control. What happens in practice remains to be seen.

Cloud computing taught practitioners how to scale resources within a single administrative domain. Edge computing requires learning how to scale in the number of administrative domains. C

Nitesh Mor is a Ph.D. candidate in computer science at the University of California, Berkeley. He is currently part of the Global Data Plane project at the Ubiquitous Swarm Lab at UC Berkeley, where he focuses on secure Internetwide infrastructures for data storage and communication.

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

Theory, Systems, and Applications

Declarative Logic Programming

Logic Programming (LP) is at the nexus of knowledge representation, AI, mathematical logic, databases, and programming languages. It allows programming to be more declarative, by specifying “what” to do instead of “how” to do it. This field is fascinating and intellectually stimulating due to the fundamental interplay among theory, systems, and applications brought about by logic.

The goal of this book is to help fill in the void in the literature with state-of-the-art surveys on key aspects of LP. Much attention was paid to making these surveys accessible to researchers, practitioners, and graduate students alike.

Edited by Michael Kifer & Yanhong Annie Liu

ISBN: 978-1-970001-96-9

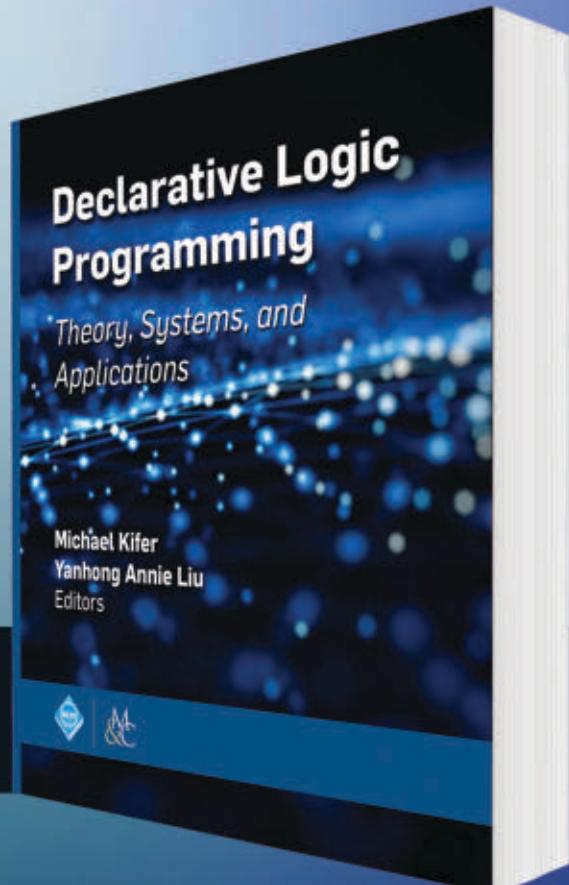
DOI: 10.1145/3191315

<http://books.acm.org>

<http://www.morganclaypoolpublishers.com/acm>



ACM BOOKS



contributed articles

DOI:10.1145/3274277

Work in finance, marketing, human resources, and operations increasingly relies on analytics—with more to come.

BY VIJAY KHATRI AND BINNY M. SAMUEL

Analytics for Managerial Work

A 2014 IDC report predicted that by 2020, the digital universe—the data we create and copy annually—will reach 44 zettabytes, or 44 trillion gigabytes.¹⁰ With the explosive growth in organizational data, there is increasing emphasis on analytics that can be used to uncover the “hidden potential” of data. A 2014 Society for Information Management survey found analytics/business intelligence to be #1 among the top 15 most significant IT investments in the prior five years.¹² It is not surprising that business analytics is increasingly central to managerial decision making within business functions: finance, marketing, human resources, and operations. For example, cash-flow analytics, shareholder-value analytics, and profit/revenue analytics

are increasingly important aspects of the finance function. A 2017 survey of chief marketing officers found companies spend 6.7% of their marketing budgets on analytics and expect to spend 11.1% over the next three years.¹⁶ A 2017 Deloitte survey of HR managers found over 71% of the surveyed companies see people analytics as a high priority.³ Analytics is increasingly used in operations management for demand forecasting, inventory optimization, spare parts optimization, warranty management, and predictive asset maintenance. Acknowledging extreme deficiency of data literacy among today’s managers, by 2020, 80% of organizations will embark on data-literacy initiatives.¹⁵

As analytics becomes central to decisions across finance-, marketing-, HR-, and operations-related work, it is also increasingly viewed as central to current and future continuous learning efforts within organizations. In the context of analytics-related continuous learning within different functions, managers need to better understand the trends in how different types of analytics applications are being and will be used for function-specific decision making. We thus focus on this research question: What current and future use of different types of analytics applications—static reports/interactive dashboards, descriptive analytics, predictive analytics, prescriptive analytics, and big data analytics—can help support different dimensions of managerial work—planning, implementing, and control-

» key insights

- A variety of analytics applications are needed to support the various dimensions of managerial work in four business functions: finance, marketing, human resources, and operations.
- The future use of analytics in these functions will employ increasingly sophisticated types of analytics applications.
- To help business managers derive value from the data in the digital economy, analytics preparedness, as well as the design of data-literacy programs, will have to be function- and work-specific.



ling—in four business functions? In 2015, we conducted a survey of 197 mid-level U.S.-based managers in finance (49), marketing (50), HR (49), and operations (49) functions on their current and future use of various analytics applications.

We found the current and future uses of various analytics applications for the four functions are different. Following our survey findings, we suggest analytics preparedness, as well as design of data-literacy programs within those functions, will need to match the use of each function. Practitioners can also use our findings to benchmark their respective organizations' current and future use of analytics applications by business functions.

The rest of this article is organized as follows: We first characterize different types of analytics applications in the four functions, then describe three dimensions of managerial work. We present our findings on the current, or 2015, and future use, or five years out, or 2020, of different types of analytics

applications to support the three dimensions. We then compare the current and future use of analytics separately in finance, marketing, HR, and operations functions, respectively.

Analytics Applications

Analytics applications are often characterized as static reports/interactive dashboards, descriptive analytics, predictive analytics, prescriptive analytics, and big data analytics. Static reports/interactive dashboards refer to database queries shown as reports and dashboards to users. Descriptive analytics employ summary statistics (such as mean, mode, and median) to characterize data and generate insights. Predictive analytics—using such tools as SAS Enterprise Miner, SPSS Modeler, and R—centers on data mining and machine learning techniques to better understand what will happen in the future based on historical data. Prescriptive analytics—simulation or optimization using such tools as @Risk and Crystal

Ball—seeks to determine what should be done in the future. Big data analytics—the use of various types of data involving the Hadoop/NoSQL ecosystem—concentrates on uncovering hidden patterns and understanding previously unknown correlations, market trends, customer preferences, and other useful information from unstructured (non-tabular) data.

We first highlight the use of analytics in four functions—finance, marketing, HR, and operations—then describe the different dimensions of managerial work.

In finance, marketing, HR, and operations functions. Finance and accounting functions are the custodians and curators of financial data. As part of the accounting/finance function, collectively referred to as the “finance function” here, planning and budgeting set targets for revenue, expenditure, and cash generation, usually relying on spreadsheets.¹¹ The objective of financial close reporting is to produce financial statements in “board book”

forms and analyze financial targets. Finally, financial forecasting and modeling rely on advanced analytics, as in, say, profit-optimization capabilities that can help gauge profitability of different strategies. A 2011 research article from <http://www.cfo.com/> reported deficiencies in current uses of analytics in finance, with approximately half of 231 companies surveyed in the U.S. reporting being less than “very effective” at incorporating information for strategic and operational decision making.

The marketing function has a long history of systematic use of data; for example, A.C. Nielsen measured product sales as early as the 1930s, geo-demographic data has been used since the 1970s, scanner panel data emerged in the 1980s, CRM software systems have been used since the 1990s, and user-generated content (such as online product reviews and blogs) in the 2000s produced large volumes of data.¹⁹ In 2004, Facebook pioneered an era of social network data, while smartphones with GPS capabilities (in 2007) prompted a flood of consumer-location data. With granular data, the marketing function increasingly employs analytics to develop and maintain customer relationships, personalize products and services, and automate marketing processes in real time, as through, say, recommendation systems and search marketing. More recently, images have been analyzed to classify facial features of models in advertisements.²⁰

The HR function is primarily responsible for managing and rationalizing employment relationships through talent management¹ and designing and overseeing performance appraisals.² Major companies (such as Capital One and Dow Chemical)

have used sophisticated simulation to better understand what will be required in terms of talent.¹ Created during World War I, the “merit rating” system—tracking past performance through numerical scores—was developed to identify poor performers and used by 60% of U.S. companies by the start of World War II, a number that was closer to 90% in the 1960s.² HR systems have traditionally used information concerning workers who are employed (as well as those not hired), hours worked, pay collected, and worker performance. More recent examples of HR analytics include workforce forecasting, human capital-investment analysis, and talent-value modeling.⁴

Operational priorities are often defined in terms of cost (such as productivity, capacity utilization, and inventory reduction), quality (such as reliability, durability, and serviceability), delivery time, and flexibility.¹⁸ Analytics is widely employed in the operations function; for example, anomaly analysis and proactive notifications can help avoid service outages; quick search across structured and unstructured data can help improve mean time to repair; supply-chain network optimization can help identify supply chain bottlenecks and support route and truckload optimization; and demand planning can help analyze customer segments in terms of channels, brands, and products down to the stock-keeping-unit level to develop models that shape demand and impact revenue.

In summary, various analytics applications are finding increasing use in a variety of business functions. In order to understand the current and future use of analytics in these functions, the next section defines managerial work in each function consistently.

In managerial work. Managerial work in business functions is often conceived as being multi-dimensional.^{6,7} One characterization of managerial work differentiates among planning, organizing, and controlling.⁷ Preventing managers from “uncritically extending the present trends into the future,” and “planning”—the first dimension—refers to integrating what the business is, what it will be, and what it should be. The “implementing” dimension refers to day-to-day organizing and daily “programming” activities. Over half of a manager’s time is spent on this dimension, making it central to managerial work.^{8,13} Finally, the process of work needs to be controlled, implying controls need to be embedded with respect to such aspects of performance as quality and efficiency; this dimension is often referred to as “controlling.”

Survey Methodology

Within each business function, we identified corresponding managerial tasks in three dimensions of managerial work: planning, implementing, and controlling. For example, to “develop product and market forecasts” is a planning-related task in the marketing function. We asked one manager in each function—finance, marketing, HR, and operations—to validate the managerial tasks for comprehensiveness. All managerial tasks we identified in the four functions are outlined in the online appendix “Managerial Work in Four Functions,” dl.acm.org/citation.cfm?doid=3274277&picked=formats.

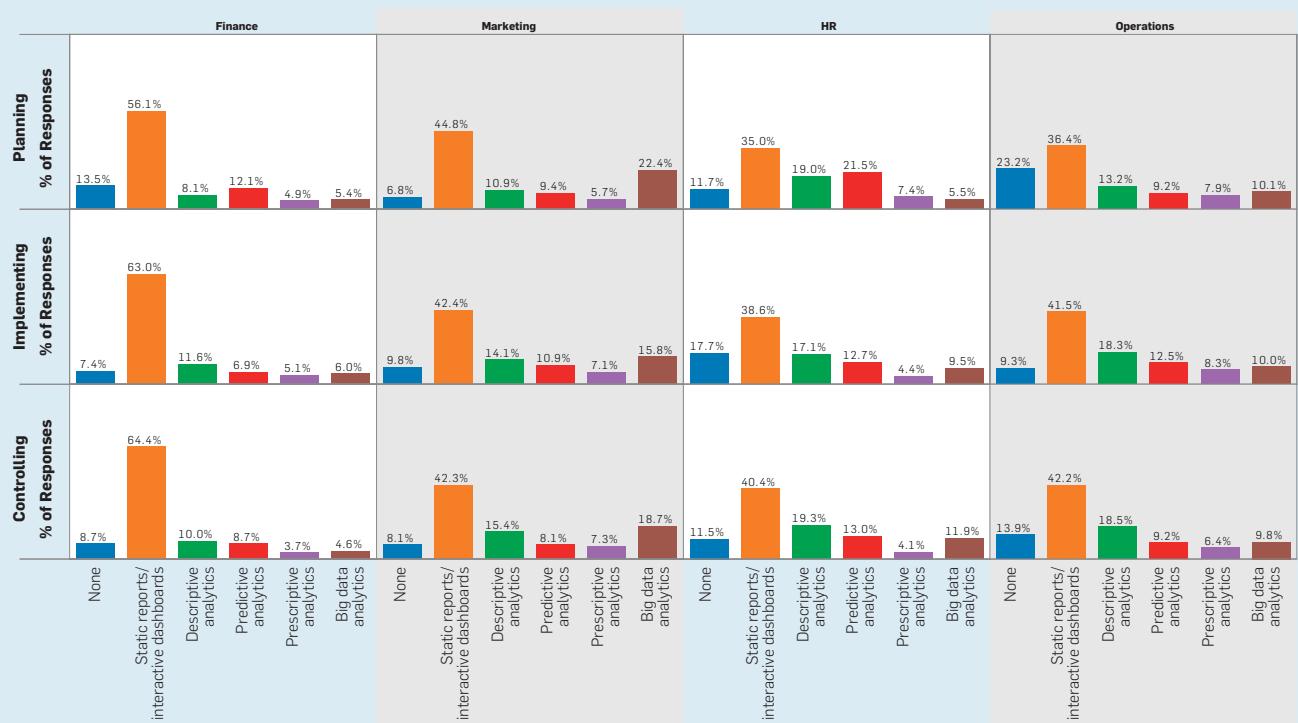
For each managerial task, the survey respondent—a finance, marketing, HR, or operations manager—could choose one or more of the following types of analytics applications currently in use for conducting that task: none, static reports/interactive dashboards, descriptive analytics, predictive analytics, prescriptive analytics, and big data analytics. None indicated analytics was not used for that managerial task. They also indicated their future use of the various analytics applications for each task.

A total of 197 individuals responded to the survey from the four functions; 49 HR, 49 finance, 49 operations, and 50 marketing. Respondents had, on average, 7.9 years of work experience in their current position and 15.3 years

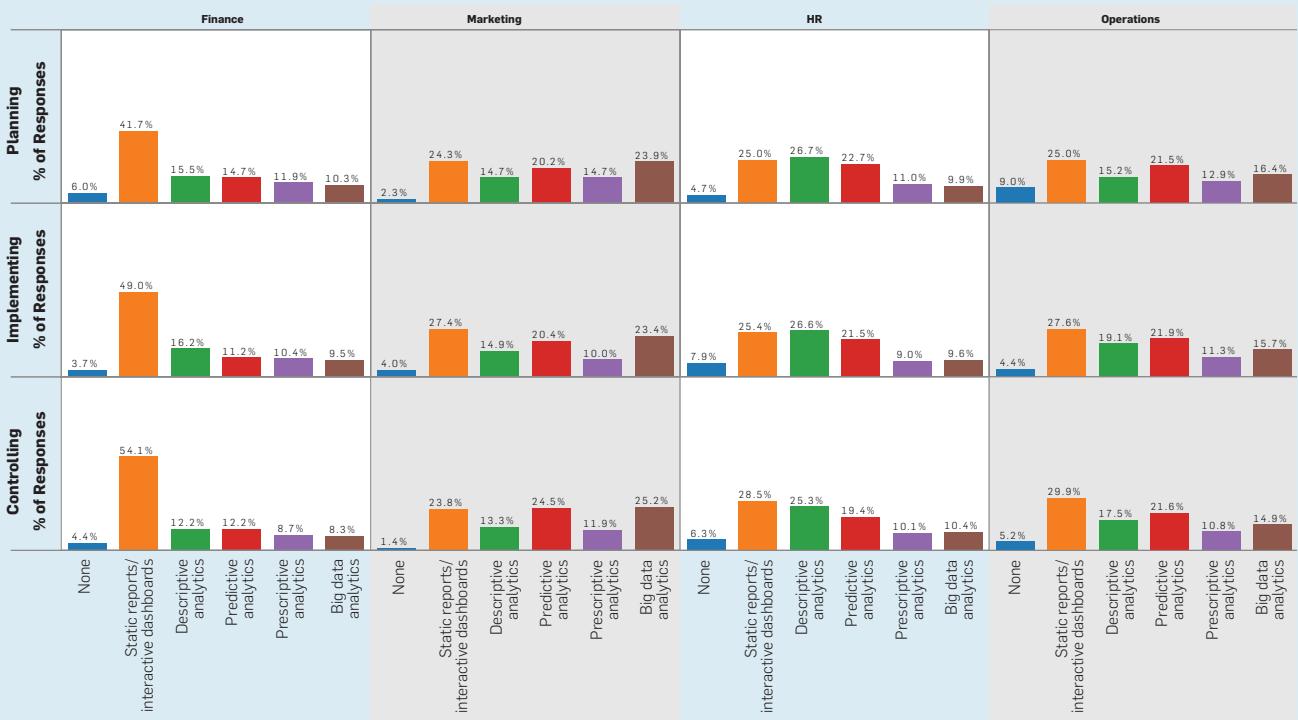
Table 1. Survey respondents ranked by organization revenue.

Organization Classification	Revenue	Number of Respondents	Percent of Respondents
Large-size organization	More than \$10 billion	34	36%
	\$5 billion–\$10 billion	10	
	\$1 billion–\$5 billion	26	
Medium-size organization	\$500 million–\$1 billion	17	38%
	\$100 million–\$500 million	21	
	\$50 million–\$100 million	15	
	\$10 million–50 million	22	
Small-size organization	Less than \$10 million	43	22%
	Do not know	9	5%

Current and future use of analytics applications for managerial work.



(a) Current use of analytics applications



(b) Future use of analytics applications

Table 2. Current vs. future use of analytics applications in finance management.

Type of Analytics Applications		Current	Future	Difference
Planning	None	13.5%	6.0%	-7.5**
	Static reports/interactive dashboards	56.1%	41.7%	-14.4**
	Descriptive analytics	8.1%	15.5%	7.4**
	Predictive analytics	12.1%	14.7%	2.6
	Prescriptive analytics	4.9%	11.9%	7.0**
	Big data analytics	5.4%	10.3%	4.9**
Implementing	None	7.4%	3.7%	-3.7
	Static reports/interactive dashboards	63.0%	49.0%	-14.0*
	Descriptive analytics	11.6%	16.2%	4.6*
	Predictive analytics	6.9%	11.2%	4.3**
	Prescriptive analytics	5.1%	10.4%	5.3**
	Big data analytics	6.0%	9.5%	3.5*
Controlling	None	8.7%	4.4%	-4.3*
	Static reports/interactive dashboards	64.4%	54.1%	-10.3*
	Descriptive analytics	10.0%	12.2%	2.2
	Predictive analytics	8.7%	12.2%	3.5
	Prescriptive analytics	3.7%	8.7%	5.0**
	Big data analytics	4.6%	8.3%	3.7*

Note

* indicates significant Chi-square difference at the p ≤ .10 alpha level

** indicates significant Chi-square difference at the p ≤ .05 alpha level

use of various analytics applications for three dimensions of managerial work—planning, implementing, and controlling—then compare the current and future use of various analytics applications for each function separately.

Our overall results are outlined in the figure here as different dimensions of managerial work along the y-axis—planning, implementing, and controlling—and business functions—finance, marketing, HR, and operations—along the x-axis. For planning-related managerial work, static reports/interactive dashboards are currently, as in Figure 1a, the most commonly used analytics applications. The current use of static reports/interactive dashboards applications was 56.1%, 44.8%, 35.0%, and 36.4% for planning-related work in finance, marketing, HR, and operations, respectively. On the other hand, the current least-used analytics application was prescriptive analytics—4.9%, 5.7%, and 7.9% for planning-related work in finance, marketing and operations functions, respectively—with the exception of the HR function where the least-used analytics application for planning-related work was big data analytics (5.5%). With respect to current implementing- and controlling-related managerial work, static reports/interactive dashboards were again the most frequently used, and prescriptive analytics was least-frequently used. Static reports/interactive dashboards and prescriptive analytics are generally the most- and least-frequently used, respectively, analytics applications for managerial work.

The prominence of static reports/interactive dashboards is not surprising considering it was the traditional view of analytics in organizations in the 1990s, and organizations have since then invested in various business-intelligence and reporting initiatives.¹⁷ Static reports/interactive dashboards tend to have less rigorous technical (mathematical) sophistication, requiring little training in use and adoption. In contrast, the lower current use of more advanced prescriptive analytics can potentially be attributed to the mathematical skills (such as linear programming) required to understand and use them. Among other analytics applications, the survey results suggest big data analytics is the most frequently used in the marketing func-

Table 3. Current vs. future use of analytics applications in marketing.

Type of Analytics Applications		Current	Future	Difference
Planning	None	6.8%	2.3%	-4.5**
	Static reports/interactive dashboards	44.8%	24.3%	-20.5**
	Descriptive analytics	10.9%	14.7%	3.8*
	Predictive analytics	9.4%	20.2%	10.8**
	Prescriptive analytics	5.7%	14.7%	9.0**
	Big data analytics	22.4%	23.9%	1.5
Implementing	None	9.8%	4.0%	-5.8**
	Static reports/interactive dashboards	42.4%	27.4%	-15.0**
	Descriptive analytics	14.1%	14.9%	0.8
	Predictive analytics	10.9%	20.4%	9.5**
	Prescriptive analytics	7.1%	10.0%	2.9
	Big data analytics	15.8%	23.4%	7.6**
Controlling	None	8.1%	1.4%	-6.7**
	Static reports/interactive dashboards	42.3%	23.8%	-18.5**
	Descriptive analytics	15.4%	13.3%	-2.1
	Predictive analytics	8.1%	24.5%	16.4**
	Prescriptive analytics	7.3%	11.9%	4.6*
	Big data analytics	18.7%	25.2%	6.5**

Note

* indicates significant Chi-square difference at the p ≤ .10 alpha level

** indicates significant Chi-square difference at the p ≤ .05 alpha level

of total work experience. Respondents worked in organizations that, on average, employed 900 full time equivalent employees. As reported in Table 1, the respondents were distributed among

small- (22%), medium- (38%), and large-size (36%) organizations.

Survey Findings

We first explore the current and future

tion, with 22.4%, 15.8%, and 18.7% for planning-, implementing-, and controlling-related marketing work. This is perhaps due to the need to better manage customer relationships with (big) data generated outside organizations on social media platforms and the long history of using analytics for marketing decisions.¹⁹ Predictive analytics applications are most frequently used for planning-, implementing-, and controlling-related HR work, with 21.5%, 12.7%, and 13.0%, respectively.

Comparing current and future use (see Figure 1b), it seems that, in general, static reports/interactive dashboards will still be the predominant analytics application in the near future. The exceptions include the projected use of descriptive analytics for planning- and implementing-related HR work (26.7% and 26.6%, respectively) and the use of big data analytics for controlling-related marketing work (25.2%). However, compared to current use, our respondents indicated the dominance of static reports/interactive dashboards will diminish in the future, in the range of 10.0 percentage points to 20.5 percentage points (henceforth referred to as “points”). This implies there is indeed an increasing trend toward using more advanced types of analytics applications—descriptive, predictive, prescriptive, and big data.

In the following paragraphs, for each business function we compare current and future use of different types of analytics applications with respect to a specific type of managerial work—planning, implementing, and controlling. For each such type of work, we highlight the greatest negative and positive difference in yellow and green (background), respectively, in Table 2, Table 3, Table 4, and Table 5. We excluded “None” from this consideration. Based on Likelihood Ratio Chi-square tests, we note statistically significant differences in each table. This test assumes a null hypothesis that the distribution of responses for using a particular analytics application is the same between the current and the future. An asterisk in the difference indicates there is a significant change in the number of respondents who reported their intention to use a specific analytics application in the future (five years out) vs. their current

use of the same analytics application.

In finance management. With respect to planning-related finance work, static reports/interactive dash-

boards (56.1%) and predictive analytics (12.1%) were the top two analytics applications currently used (see Table 2). In the future, static reports/interactive

Table 4. Current vs. future use of analytics applications in HR management.

Type of Analytics Applications	Current	Future	Difference	
Planning	None	11.7%	4.7%	-7.0**
	Static reports/interactive dashboards	35.0%	25.0%	-10.0*
	Descriptive analytics	19.0%	26.7%	7.7**
	Predictive analytics	21.5%	22.7%	1.2
	Prescriptive analytics	7.4%	11.0%	3.6
Implementing	Big data analytics	5.5%	9.9%	4.4*
	None	17.7%	7.9%	-9.8**
	Static reports / interactive dashboards	38.6%	25.4%	-13.2*
	Descriptive analytics	17.1%	26.6%	9.5**
	Predictive analytics	12.7%	21.5%	8.8**
Controlling	Prescriptive analytics	4.4%	9.0%	4.6**
	Big data analytics	9.5%	9.6%	0.1
	None	11.5%	6.3%	-5.2**
	Static reports/interactive dashboards	40.4%	28.5%	-11.9**
	Descriptive analytics	19.3%	25.3%	6.0**
	Predictive analytics	13.0%	19.4%	6.4**
	Prescriptive analytics	4.1%	10.1%	6.0**
	Big data analytics	11.9%	10.4%	-1.5

Note

* indicates significant Chi-square difference at the p ≤ .10 alpha level

** indicates significant Chi-square difference at the p ≤ .05 alpha level

Table 5. Current vs. future use of analytics applications in operations management.

Type of Analytics Applications	Current	Future	Difference	
Planning	None	23.2%	9.0%	-14.2**
	Static reports/interactive dashboards	36.4%	25.0%	-11.4**
	Descriptive analytics	13.2%	15.2%	2.0
	Predictive analytics	9.2%	21.5%	12.3**
	Prescriptive analytics	7.9%	12.9%	5.0**
Implementing	Big data analytics	10.1%	16.4%	6.3**
	None	9.3%	4.4%	-4.9**
	Static reports/interactive dashboards	41.5%	27.6%	-13.9**
	Descriptive analytics	18.3%	19.1%	0.8
	Predictive analytics	12.5%	21.9%	9.4**
Controlling	Prescriptive analytics	8.3%	11.3%	3.0*
	Big data analytics	10.0%	15.7%	5.7**
	None	13.9%	5.2%	-8.7**
	Static reports/interactive dashboards	42.2%	29.9%	-12.3*
	Descriptive analytics	18.5%	17.5%	-1.0
	Predictive analytics	9.2%	21.6%	12.4**
	Prescriptive analytics	6.4%	10.8%	4.4*
	Big data analytics	9.8%	14.9%	5.1*

Note

* indicates significant Chi-square difference at the p ≤ .10 alpha level

** indicates significant Chi-square difference at the p ≤ .05 alpha level

dashboards (41.7%) and descriptive analytics (15.5%) will be popular applications. Comparing current to future use of analytics applications for planning-related finance work, the respondents in general indicated a significant increase in the use of advanced analytics applications in the future—descriptive (by 7.4 points), prescriptive (by 7.0 points), and big data analytics (by 4.9 points). On the other hand, static reports/interactive dashboards are projected to see the greatest drop in use (by 14.4 points).

With respect to implementing-related finance work, static reports/interactive dashboards alone account for almost two-thirds of the current analytics applications used (63.0%). Comparing current and future use of analytics applications, prescriptive analytics has the most significant increase in use (by 5.3 points), while static reports/dashboards are expected to have the most significant decrease in use (by 14.0 points).

For controlling-related finance work, static reports/interactive dashboards were the dominant application being used (64.4%) and will continue to be dominant in the future (54.1%). In the future, our respondents indicated an overall increase in the use of different types of advanced analytics applications, though the increase is only significant for prescriptive and big data analytics, by 5.0 and 3.7 points, respectively. Much like planning- and implementing-related finance work, static reports/interactive dashboards are projected to see the greatest drop in use (by 10.3 points).

By percentage points, prescriptive analytics will experience the greatest increase in implementing- and controlling-related finance work, by 5.3 and 5.0 points, respectively, and is a close second for the greatest increase in planning-related finance work (by 7.0 points) behind descriptive analytics. Our findings suggest educating finance managers in prescriptive and descriptive analytics applications will be imperative for success in their job performance. While the finance function is quantitative by nature, prior findings recognized it trailed other functions (such as marketing, operations, and HR) in its use of analytics.⁵ More recently though, Davenport and Tay⁵ provided

an example of how Intel has begun initiatives in forecasting revenue and predicting impairments in its capital investments. Our findings suggest the finance function may be moving beyond static reports toward using advanced analytics more broadly.^{5,11}

In marketing management. With respect to planning-related marketing work, our respondents indicated current use to be the highest for static reports/interactive dashboards (44.8%) followed by big data analytics (22.4%) (see Table 3). In the future, a significantly greater share of planning-related marketing work will likely involve descriptive (14.7%), predictive (20.2%), and prescriptive analytics (14.7%), with a concomitant decrease in the use of static reports/interactive dashboards (by 20.5 points).

With respect to implementing-related marketing work, static reports/interactive dashboards are currently the most frequently used analytics application (42.4%). The dominance of this type of application will decrease by 15 points in the future. On the other hand, predictive and big data analytics will see the greatest increases in the future, with overall usage at 20.4% and 23.4%, respectively.

Controlling-related marketing work follows similar current usage trends as the other two dimensions—planning and implementing. The use of static reports/interactive dashboards will again significantly decrease in the future (by 18.5 points), and this will, it seems, be replaced by a significant increase in the use of predictive (by 16.4 points) analytics. The use of prescriptive (by 4.6 points) and big data (by 6.5 points) analytics are also projected to increase significantly.

Our findings suggest marketing work is leading the use of big data analytics applications and is expected to continue in the future. Comparing current to future use, there is a general shift toward more advanced analytics applications—predictive, prescriptive, and big data—for various tasks in the marketing function, with predictive analytics experiencing the greatest percentage points increase. This suggests data-literacy programs will need to provide marketing managers with ways to incorporate predictive analytics in their work. The marketing function

involves the greatest projected drop in the use of static reports/interactive dashboards among all the functions. This is consistent with prior research that suggests recommendations, geo-fencing, search marketing, and retargeting, or types of advanced analytics, are increasingly being embedded into the marketing function.¹⁹

In HR management. With respect to planning-related HR work, static reports/interactive dashboards are currently the most used (35.0%), with predictive analytics generally being the second most used (21.5%) (see Table 4). In the future, descriptive analytics will experience a significant increase (by 7.7 points), matching a similar decrease in static reports/interactive dashboards (by 10 points). We also found the use of big data analytics is expected to increase significantly (by 4.4 points).

Implementing-related HR work has a similar current usage pattern as planning-related HR work in terms of leading current and future analytics applications. In this dimension of HR work, the use of static reports/interactive dashboards will decrease significantly (by 13.2 points), along with a significant increase in the use of descriptive analytics (by 9.5 points). Additionally, the projected use of predictive analytics will significantly increase in the future (by 8.8 points) to 21.5%. While prescriptive analytics will generally continue to be the least frequently used analytics application in the future, its use will nonetheless significantly increase in the future (by 4.6 points).

Similar to implementing-related HR work, for controlling-related HR work, the data indicated similar significant increases in descriptive, predictive, and prescriptive analytics (by approximately 6 points) for controlling-related HR work and a concomitant decrease in projected use of static reports/interactive dashboards (by 11.9 points).

Overall, descriptive analytics will play a greater role in HR work, as it is expected to experience the greatest point increase for planning and implementing and be tied for the second greatest increase for controlling. This makes it an important analytics application for future data-literacy programs in HR. Given that, in 2014, 12% of U.S.-based companies, includ-

ing Adobe, Juniper Systems, Dell, and Microsoft, had dropped traditional one-time annual employee appraisals, the increased use of descriptive analytics for more frequent performance appraisals is not surprising.² Our respondents also indicated increasing use of advanced analytics applications for HR work, or predictive and prescriptive analytics, and these increases were significant for implementing and controlling dimensions. Predictive analytics applications include predicting an employee's or a job candidate's personality and future work performance based on their Facebook profile.¹⁴

In operations management. Static reports/interactive dashboards are the most used current analytics application for planning-related operations work (36.4%) (see Table 5). In the future, such use will decrease significantly, by 11.4 points to 25%. Predictive analytics will have a significant and almost-matching increase (by 12.3 points) and be the second most used application for planning-related operations work (21.5%). Prescriptive and big data analytics will also increase significantly, by 5.0 and 6.3 points, respectively.

With respect to implementing-related operations work, static reports/interactive dashboards were the most used analytics application (41.5%), but their use is expected to decrease significantly in the future (by 13.9 points). This decrease will be replaced primarily by a significant increase in the use of predictive analytics (by 9.4 points); additionally, future use of prescriptive and big data analytics will also increase significantly, by 3.0 and 5.7 points, respectively.

In controlling-related operations work, while static reports/interactive dashboards are the most (42.2%) used today, this use is projected to decrease significantly by 12.3 points to 29.9%. Predictive analytics in controlling-related operations work will experience a significant (and the greatest) percentage-point increase in use (12.4 points) to 21.6%, and prescriptive and big data analytics will both increase significantly as well, by 4.4 and 5.1 points, respectively.

We note the operations function had a similar pattern of current and future analytics application use and

Given that analytics is increasingly being embedded in managerial work, “one size fits all” cannot be employed for domain-specific decision making.

pattern of change across all three dimensions of work. Predictive analytics is poised for the greatest share increase, and data-literacy programs will need to support this type of analytics application for operations work in the near future. The increased use of predictive analytics in operations management is potentially due to its greater use in product life-cycle management, asset management, inventory management, and service management. For example, optimal maintenance schedules can be set to reduce downtime, and alerts can be received about any imminent failure. As another example, Volvo conducts predictive, machine learning-driven analytics across petabyte-scale datasets to discern breakdown and failure rates.^a

Conclusion

While widely recognized that analytics increasingly influences work in various business functions, no prior research has methodically examined how various types of analytics applications support different dimensions of managerial work in four business functions. Given that analytics is increasingly being embedded in managerial work, “one size fits all” cannot be employed for domain-specific decision making.¹⁵ Even *within* a business function, the prevalence of analytics applications might differ by the associated dimensions of work—planning, implementing, and controlling. Future data-literacy programs will thus need to be designed differently for each business function.

Our survey included almost 200 U.S.-based business managers to provide a nuanced perspective on the current and future use of different types of analytics applications for supporting managerial work in four business functions within an organization. Here, we offer some concluding comments regarding our findings. First, compared to current use, we found indications of significant increases in future use of advanced analytics applications. Consistent with prior studies, the survey findings suggest the diversified use of

^a <https://www.forbes.com/sites/bernard-marr/2016/07/18/how-the-connected-car-is-forcing-volvo-to-rethink-its-data-strategy/>

analytics applications is on the rise within all four functions. As different types of data become available for decision making (such as time-series scanner panel data, text-based user-generated content, social network data, and consumer location data¹⁹) our findings suggest that, in the future, managers will need to be engaged in different types of analytics applications as a core aspect of their responsibility.

Second, in general, we found static reports/interactive dashboards are and will continue to be a frequently used analytics application for all four functions. Our survey findings suggest managers need to continue to develop competency in their use of static reports/interactive dashboards, and business intelligence initiatives. However, their use will decrease significantly—in the range of 10 percentage points to 20.5 percentage points. This is consistent with a Gartner report that found spending on traditional business intelligence has been decreasing since 2015, with concomitant broad and pervasive deployment of self-service analytics.⁹ This decrease is happening as software for advanced analytics applications is becoming easier to use, and, with increased data literacy, business functions are starting to discover how more-sophisticated analytics applications can help managers accomplish their respective work. For example, dedicated applications that can predict financial riskiness of individuals based on their mouse movements while completing loan applications are now being used by the finance function.^b

Third, we found the functions will differ as to which will be most used in the future. We found that the managerial work in the finance function will continue to rely on static reports/interactive dashboards the most, with projected future use at 41.7%, 49%, and 54.1% for planning-, implementing-, and controlling-related work, respectively. On the other hand, reliance on static reports/interactive dashboards is relatively lower for marketing (between 23.8% and 27.4%), HR (between 25% and 28.5%), and operations (between

25% and 29.9%) functions. HR will have static reports/interactive dashboards, descriptive analytics, and prescriptive analytics applications as the top three analytics applications. Marketing and operations can expect similar future use patterns, with static reports/interactive dashboards, predictive analytics, and big data analytics applications as their top-three analytics applications. These patterns suggest data-literacy programs and associated analytics investments will need to be function- and work-specific in the future.

Fourth, we found the functions are changing at different rates in terms of their expected future use of analytics applications for different dimensions of work. The finance function will see the greatest point increases in prescriptive analytics for implementing-related work. On the other hand, the HR function will see the greatest point increase in descriptive analytics for all dimensions of work. Both marketing and operations functions will experience the greatest point increases in predictive analytics across all dimensions of work. Our data indicates across all dimensions of managerial work, the finance function lags in today's use of an analytics application nine times, while that lag for HR, marketing, and operations occurs four, one, and one time, respectively. However, in the future, the finance function will lag six times, while HR, marketing, and operations will lag four, five, and zero times, respectively. These results suggest some functions are more quick to embrace the capabilities of analytics to help them accomplish their work.

Finally, these survey results provide a benchmark for different functions in their current and future use of different types of analytics applications. Our findings can be used as a starting point for discussion within organizations about their current and anticipated future use of analytics to support different types of managerial work.

Acknowledgments

We thank Andrew A. Chien, Robert D. Austin, and the anonymous reviewers for providing very helpful comments that improved the quality of this article.

References

1. Cappelli, P. Talent management for the 21st century. *Harvard Business Review* 86, 3 (2008), 1–9.
2. Cappelli, P. and Tavis, A. The performance management revolution. *Harvard Business Review* 94, 10 (2016), 58–67.
3. Collins, L., Fineman, D.R., and Tsuchida, A. People analytics: Recalculating the route. *Deloitte Insights* (Feb. 28, 2017), <https://www2.deloitte.com/insights/us/en/focus/human-capital-trends/2017/people-analytics-in-hr.html>
4. Davenport, T.H., Harris, J., and Shapiro, J. Competing on talent analytics. *Harvard Business Review* 88, 10 (2010), 52–58.
5. Davenport, T.H. and Tay, A. Finance must ramp up role as analytics leader. *Argyle Executive Forum CFO*, Aug. 26, 2016, <http://www2.cfo.com/analytics/2016/08/finance-must-ramp-role-analytics-leader/>
6. Drucker, P. *Management: Tasks, Responsibilities, Practices*. Harper & Row, Heinemann, London, U.K., 1974.
7. Fayol, H. *General and Industrial Management*. Pitman, London, U.K., 1949.
8. Horne, J.H. and Lupton, T. The work activities of 'middle' managers: An exploratory study. *Journal of Management Studies* 2, 1 (Feb. 1965), 14–33.
9. Howson, C., Sallam, R.L., Richardson, J.L., Tapadinhas, J., Idoine, C. J., and Woodward, A. *Magic Quadrant for Analytics and Business Intelligence Platforms*. Gartner, Stamford, CT, Feb. 26, 2018; <https://www.gartner.com/document/3861464>
10. IDC. *Executive Summary Data Growth, Business Opportunities, and the IT Imperatives*. Framingham, MA, Apr. 2014; <https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm>
11. Iervolino, C. and Van Decker, J.E. *The Future of Financial Planning and Analysis*. Gartner, Stamford, CT, Nov. 1, 2017; <https://www.gartner.com/document/3823072>
12. Kappelman, L., McLean, E., Johnson, V., and Gerhart, N. The 2014 SIM IT key issues and trends study. *MIS Quarterly Executive* 13, 4 (2014), 237–263.
13. Kelly, J. The study of executive behaviour by activity sampling. *Human Relations* 17, 3 (Aug. 1964), 277–287.
14. Kluepmer, D.H., Rosen, P.A., and Mossholder, K.W. Social networking websites, personality ratings, and the organizational context: More than meets the eye? *Journal of Applied Social Psychology* 42, 5 (May 2012), 1143–1172.
15. Logan, V.A. *Information as a Second Language: Enabling Data Literacy for Digital Society*. Gartner, Stamford, CT, Sept. 21, 2018; <https://www.gartner.com/document/3890564>
16. Moorman, C. *The CMO Survey*. World Market Watch LLC, Chapel Hill, NC, Feb. 2017; <https://cmosurvey.org/results/february-2017/>
17. Richardson, J.L., Tapadinhas, J., Howson, C., Sallam, R.L., Idoine, C.J., and Woodward, A. *Critical Capabilities for Analytics and Business Intelligence Platforms*. Gartner, Stamford, CT, May 7, 2018; <https://www.gartner.com/document/3874285>
18. Ward, P.T., McCreery, J.K., Ritzman, L.P., and Sharma, D. Competitive priorities in operations management. *Decision Sciences* 29, 4 (Sept. 1998), 1035–1046.
19. Wedel, M. and Kannan, R.K. Marketing analytics for data-rich environments. *Journal of Marketing* 80, 6 (Nov. 2016), 97–121.
20. Xiao, L. and Ding, M. Just the faces: Exploring the effects of facial features in print advertising. *Marketing Science* 33, 3 (Feb. 2014), 338–352.

Vijay Khatri (vkhatri@indiana.edu) is a professor and Arthur M. Weimer faculty fellow in the Operations and Decision Technologies Department of Indiana University's Kelley School of Business, Bloomington, IN, USA.

Binny M. Samuel (samuelby@uc.edu) is an assistant professor in the Operations, Business Analytics, and Information Systems Department of the Lindner College of Business at the University of Cincinnati, Cincinnati, OH, USA.

^b <https://www.wsj.com/articles/your-moods-change-the-way-you-move-your-mouse-1452268410>

Introducing ACM Transactions on Internet of Things (TIOT)

A new journal from ACM publishing novel research contributions and experience reports in domains whose synergy and interrelations enable the IoT vision

Now Accepting Submissions

ACM Transactions on Internet of Things (TIOT) publishes novel research contributions and experience reports in several research domains whose synergy and interrelations enable the IoT vision. TIOT focuses on system designs, end-to-end architectures, and enabling technologies, and on publishing results and insights corroborated by a strong experimental component.

Submissions are expected to provide experimental evidence of their effectiveness in realistic scenarios and the related datasets. The submission of purely theoretical or speculative papers is discouraged, and so is the use of simulation as the sole form of experimental validation.

Experience reports about the use or adaptation of known systems and techniques in real-world applications are equally welcome, as these studies elicit precious insights for researchers and practitioners alike. For this type of submissions, the depth, rigor, and realism of the experimental component is key, along with the analysis and expected impact of the lessons learned.

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



Association for
Computing Machinery

Advancing Computing as a Science & Profession



review articles

DOI:10.1145/3231589

Advances in neurotechnologies are reigniting opportunities to bring neural computation insights into broader computing applications.

BY JAMES B. AIMONE

Neural Algorithms and Computing Beyond Moore's Law

THE IMPENDING DEMISE of Moore's Law has begun to broadly impact the computing research community.³⁸ Moore's Law has driven the computing industry for many decades, with nearly every aspect of society benefiting from the advance of improved computing processors, sensors, and controllers. Behind these products has been a considerable research industry, with billions of dollars invested in fields ranging from computer science to electrical engineering. Fundamentally, however, the exponential growth in computing described by Moore's Law was driven by advances in materials science.^{30,37} From the start, the power of the computer has been limited by the density of transistors. Progressive advances in how to manipulate silicon through advancing lithography methods and new design tools have kept advancing

computing in spite of perceived limitations of the dominant fabrication processes of the time.³⁷

There is strong evidence that this time is indeed different, and Moore's Law is soon to be over for good.^{3,38} Already, Dennard scaling, Moore's Law's lesser known but equally important parallel, appears to have ended.¹¹ Dennard's scaling refers to the property that the reduction of transistor size came with an equivalent reduction of required power.⁸ This has real consequences—even though Moore's Law has continued over the last decade, with feature sizes going from ~65nm to ~10nm; the ability to speed up processors for a constant power cost has stopped. Today's common CPUs are limited to about 4GHz due to heat generation, which is roughly the same as they were 10 years ago. While Moore's Law enables more CPU cores on a chip (and has enabled high power systems such as GPUs to continue advancing), there is increasing appreciation that feature sizes cannot fall much further, with perhaps two or three further generations remaining prior to ending.

Multiple solutions have been presented for technological extension of Moore's Law,^{3,33,38,39} but there are two main challenges that must be addressed. For the first time, it is not immediately evident that future materials

» key insights

- While Moore's Law is slowing down, neuroscience is experiencing a revolution, with technology enabling scientists to have more insights into the brain's behavior than ever before and thus positioning the neuroscience field to provide a long-term source of inspiration for novel computing solutions.
- Extending the reach of brain-inspiration into computing will not only make current AI methods better, but looking beyond the brain's sensory systems can also expand the reach of AI into new applications.
- Realizing the full potential of brain-inspired computing requires increased collaborations and sharing of knowledge between the neuroscience, computer science, and neuromorphic hardware communities.



will be capable of providing a long-term scaling future. While non-silicon approaches such as carbon nanotubes or superconductivity may yield some benefits, these approaches also face theoretical limits that are only slightly better than the limits CMOS is facing.³¹ Somewhat more controversial, however, is the observation that requirements for computing are changing.^{33,39} In some respects, the current limits facing computing lie beyond what the typical consumer outside of the high-performance computing community will ever require for floating point math. Data-centric computations such as graph analytics, machine learning, and searching large databases are increasingly pushing the bounds of our systems and are more relevant for a computing industry built around mobile devices and the Internet. As a result, it is reasonable to consider the ideal computer is not one that is better at more FLOPS, but rather one that is capable of providing low-power computation more appropriate for a world flush with “big data.” While speed re-

mains an important driver, other considerations—such as algorithmic capabilities—are increasingly critical.

For these reasons, neural computing has begun to gain increased attention as a post-Moore’s Law technology. In many respects, neural computing is an unusual candidate to help extend Moore’s Law. Neural computing is effectively an algorithmic and architectural change from classic numerical algorithms on von Neumann architectures, as opposed to exploiting a novel material to supplant silicon. Further, unlike quantum computation, which leverages different physics to perform computation, neural computing likely falls within the bounds of classic computing theoretical frameworks. Whereas quantum computation can point to exponential benefits on certain tasks such as Shor’s quantum algorithm for factoring numbers;³⁴ neural computing architecture’s most likely path to impact is through polynomial trade-offs between energy, space, and time. Such benefits can be explicitly

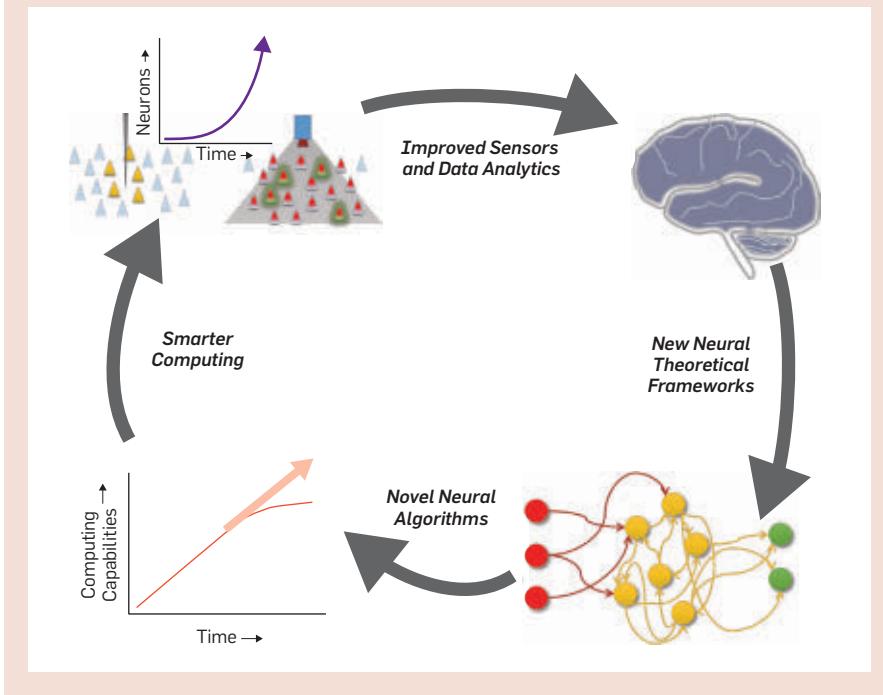
formalized and are potentially quite impactful for certain applications.¹ However, there is limited evidence that neural architectures can be more powerful on generic applications than the general purpose architectures used today.³³

The identification of neuromorphic technologies as a potential driver beyond Moore’s Law³⁹ forces the question of whether neural computing is truly a paradigm that will permit exponential scaling going forward, or rather would it represent a “one-off” gain of efficiency in some dimension, such as power efficiency? While potentially impactful, such a value proposition would not represent a long-lasting scaling capability. While to some the distinction between these two futures may appear semantic, there is a considerable difference. If neural architectures indeed represent only a one-time gain to accelerate a handful of algorithms, then it perhaps merits some consideration by specialized communities. However, if neural computation were to actually represent a scalable technology, it would justify a significant research investment from the trillion-dollar computing industry.

This article posits that new computational paradigms that leverage emerging neuroscience knowledge represent a distinctly new foundation for scaling computing technology going forward. Instead of relying on continual advances in miniaturization of devices; neural computing is positioned to benefit from long-lasting intellectual advances due to our parallel gain of knowledge of the brain’s function (Figure 1). In effect, because the materials science and chemistry of devices has been extensively optimized, we may achieve greater impact by looking to the brain for neural inspiration and hopefully achieve a continual advancement of our neural computing capabilities through algorithmic and architectural advances. Arguably, the recent successes of deep artificial neural networks (ANNs) on artificial intelligence applications is a compelling first step of this process, but the perspective offered here will contend that more extensive incorporation of insights from brain will only continue to improve our computational capabilities. This influence of neu-

Figure 1. Moore’s Law has helped initiate a potential positive feedback loop between neural data collection and improved computation.

Moore’s Law has enabled the miniaturization of sensors and improved the analytics necessary to improve neural data collection. This increased neural data has the potential to dramatically improve our ability to extract knowledge from the brain and incorporate deeper brain-derived capabilities into new algorithms and architectures; in turn furthering the advances of computing technology.



roscience can then more effectively transition to novel computational architectures and more efficient use of silicon's capabilities.

Knowledge of the Brain Is Undergoing Its Own Dramatic Scaling

Several critical efforts are underway that make such a revolutionary perspective possible. Major government funded efforts in neuroscience, such as the BRAIN Initiative in the U.S., the Human Brain Project in the European Union, and the China Brain Project, are focused on studying the brain at a systems level that they argue will maximize the computational understanding of neural circuits. Several major non-profit efforts, most notably the Allen Institute for Brain Sciences and the Howard Hughes Medical Institute Janelia Research Campus, similarly have developed large programs to systematically study neural circuits. As a specific example, the BRAIN Initiative has a guiding goal of recording a million neurons simultaneously in awake, behaving animals.⁴ Such a goal would have been unfathomable only a few years ago; however, today it increasingly appears within neuroscientists' grasp. Somewhat ironically, the advances in neuroscience sensors which allow neuroscientists to measure the activity of thousands of neurons at a time have been fueled in large part by the miniaturization of devices described by Moore's Law. It has been noted the increase in numbers of neurons recorded within a single experiment has itself undergone an exponential scaling over recent decades.³⁶

Similarly, large-scale efforts seeking to reconstruct the "connectome" of the brain are becoming more common.¹⁹ In contrast to ANNs, neural circuits are highly complex and vary considerably across brain regions and across organisms. This connectome effectively represents the graph on which biological neural computation occurs, and many neuroscientists argue that knowing this connectivity is critical for understanding the wide range of neural computations performed by the brain. While the technology to image these large-scale connectomes is increasingly available, there is a growing appreciation that challenges sur-

New computational paradigms that leverage emerging neuroscience knowledge represent a distinctly new foundation for scaling computing technology going forward.

rounding data analysis and storage are likely to become the limiting factor of neurotechnology as opposed to simply achieving higher resolution sensor technologies.^{5,12}

This rise of large-scale neuroscience efforts focused on high-throughput characterization of the brain rests on many decades of substantial progress in understanding biological neural circuits, but it is notable that neuroscience's influence on computation has been relatively minor. While neural networks and related methods have been experiencing a renaissance in recent years, the advances that led to deep learning did not derive from novel insights about neurobiological processing, but rather from a few key algorithmic advances and the availability of large-volumes of training data high-performance computing platforms such as GPUs.²³

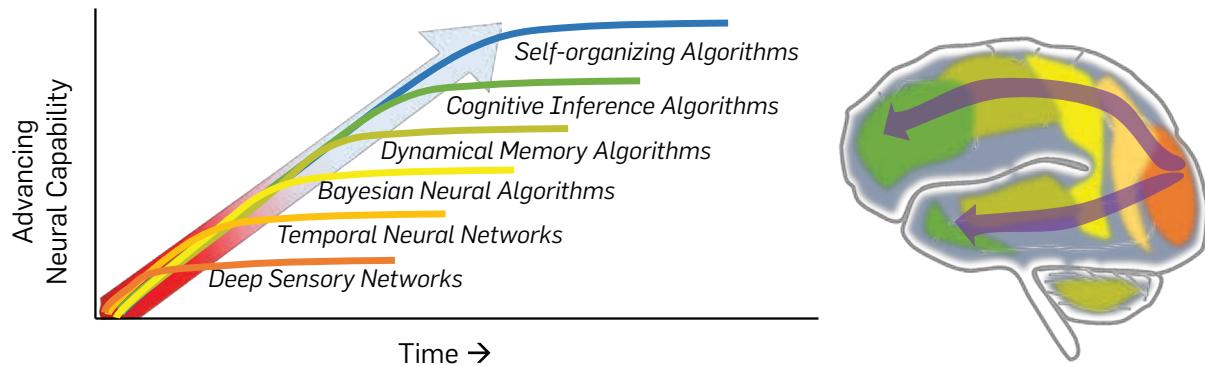
While advances in neuroscience are not responsible for the recent successes in machine learning, there are reasons that it will be more important to look to the brain going forward. For example, the brain may offer novel computational mechanisms that enable the machine learning field to impact domains that still require human intervention, in the same sense that Moore's Law benefited from disruptive shifts in materials science. Two such areas are the requirements of current deep learning techniques for high-quality training data and the capabilities targeted by machine learning applications. Of most immediate concern is the data requirement of machine learning methods. While large volumes of data are increasingly common in many applications, obtaining high-quality data—defined by both well calibrated sensors and effective annotations—is often incredibly expensive and time consuming. As a result, the ability for deep learning-related methods to impact domains with inappropriately structured data has been limited, even in domains where this is relatively straightforward for human operators.

More efficient use of data is an area of intensive machine learning research today, and has seen some recent improvements with regularization techniques such as "dropout"³⁵ and generative adversarial networks, or GANs,

Figure 2. The continued scaling of neural computing need not rely on improved materials, but rather can be achieved by looking elsewhere within the brain.

Today, we are exploiting advances of conventional ANNs at large scale, but there are already trends toward more temporal based neural networks such as long short-term memory. We are poised to benefit from a series of these technological advances, bringing neural algorithms closer to the more sophisticated computational potential of the brain.

Algorithm Class	Current Algorithms	Inspiration	Application
Deep Vision Processing	Deep Convolutional Networks (VGG, AlexNet, GoogleNet), HMax, Neocognitron	Hierarchy of sensory nuclei and early sensory cortices	Static feature extraction (e.g., images) and pattern classification
Temporal Neural Networks	Deep Recurrent Networks (e.g., long short-term memory), Hopfield Networks	Local recurrence of most biological neural circuits, especially higher sensory cortices	Dynamic feature extraction (e.g., videos, audio) and classification
Bayesian Neural Algorithms	Predictive Coding, Hierarchical Temporal Memory, Recursive Cortical Networks	Substantial reciprocal feedback between "higher" and "lower" sensory cortices	Inference across spatial and temporal scales
Dynamical Memory and Control Algorithms	Liquid State Machines, Echo State Networks, Neural Engineering Framework	Continual dynamics of hippocampus, cerebellum, and prefrontal and motor cortices	Online learning content-addressable memory and adaptive motor control
Cognitive Inference Algorithms	Reinforcement learning (e.g., Deep Q-learning) Neural Turing Machines	Integration of multiple modalities and memory into prefrontal cortex, which provides top-down influence on sensory processing	Context and experience dependent information processing and decision making
Self-organizing Algorithms	Neurogenesis Deep Learning	Initial development and continuous refinement of neural circuits to specific input and outputs	Automated neural algorithm development for unknown input and output transformations



that help utilize poorly labeled data,¹⁴ but there are many reasons to believe that the brain's approach to maximizing the utility of observed data in both developmental and adult learning is a notable area where brain-inspiration can dramatically improve computing. More broadly, it is useful to consider neuroscience's impact on computing capabilities. In general, machine learning has focused primarily on tasks most associated with sensory processing in the brain. The increased knowledge of neural circuits of non-sensory regions, such as the hippocampus, pre-frontal cortex, and striatum, may well provide opportunities for radically different

approaches to algorithms that provide computational intelligence.

A Potential Timeline of Brain-Inspired Capabilities in Computation

Here, I describe one outlook for neural algorithm "scaling," wherein the community benefits from the development of progressively more advanced brain-like capabilities in algorithms. This work comes from the perspective that the rapid increase in available experimental data is a trend that is unlikely to end soon. While the BRAIN Initiative goal of simultaneously recording one million neurons may ap-

pear impressive, that number of neurons is only a tiny fraction of a rodent cortex; and the diversity of neural regions and complex behaviors suggests a plethora of algorithms wait to be defined. A major indicator of this progress will be potential developments in theoretical neuroscience. Neuroscience has long been a field where the ability to collect data has constrained the development of robust neural theories, and it is a growing hope that the ability to measure large populations of neurons simultaneously will inspire the development of more advanced neural theories that previously would have been dismissed as

non-falsifiable due to our inability to perform the requisite experiments in the brain.^{7,40}

Figure 2 illustrates one potential path by which more sophisticated neural algorithms could emerge going forward. Because of advances in neuroscience, it is reasonable to expect the current deep learning revolution could be followed by complementary revolutions in other cognitive domains. Each of these novel capabilities will build on its predecessors—it is unlikely that any of these algorithms will ever go away, but it will continue to be used as an input substrate for more sophisticated cognitive functions. Indeed, in most of the following examples, there is currently a deep learning approach to achieving that capability (noted in the second column). Importantly, this description avoids an explicit judgment between the value of neural-inspired and neural-plausible representations; the neuroscience community has long seen value in representing cognitive function at different levels of neural fidelity. Of course, for computing applications, due to distinct goals from biological brains, it is likely that some level of abstraction will often outperform algorithms relying on mimicry; however, for theoretical development, it will likely be more effective for researchers to represent neural computation in a more biologically plausible fashion.

1. Feed-forward sensory processing.

The use of neural networks to computationally solve tasks associated with human vision and other sensory systems is not a new technology. While there have been a few key theoretical advances, at their core deep learning networks, such as deep convolutional networks, are not fundamentally different from ideas being developed in the 1980s and 1990s. As mentioned earlier, the success of deep networks in the past decade has been driven in large part due to the availability of sufficiently rich datasets as well as the recognition that modern computing technology, such as GPUs, are effective at training at large scale. In many ways, the advances that have enabled deep learning have simply allowed ANNs to realize the potential that the connectionist cognitive science community has been predicting for several decades.

From a neuroscience perspective, deep learning's success is both promising and limited. The pattern classification function that deep networks excel at is only a very narrow example of cognitive functionality, albeit one that is quite important. The inspiration it takes from the brain is quite restricted as well. Deep networks are arguably inspired by neuroscience that dates to the 1950s and 1960s, with the recognition by neuroscientists like Vernon Mountcastle, David Hubel, and Torsten Wiesel that early sensory cortex is modular, hierarchical, and has representations that start simple in early layers and become progressively more complex. While these are critical findings that have also helped frame cortical research for decades, the neuroscience community has built on these findings in many ways that have yet to be integrated into machine learning. One such example, described here, is the importance of time.

2. Temporal neural networks. We appear to be at a transition point in the neural algorithm community. Today, much of the research around neural algorithms is focused on extending methods derived from deep learning to operate with temporal components. These methods, including techniques such as long short-term memory, are quickly beginning to surpass state of the art on more time-dependent tasks such as audio processing.²³ Similar to more conventional deep networks, many of these time-based methods leverage relatively old ideas in the ANN community around using network recurrence and local feedback to represent time.

While this use of time arising from local feedback is already proving powerful, it is a limited implementation of the temporal complexity within the brain. Local circuits are incredibly complex; often numerically dominating inputs and outputs to a region and consisting of many distinct neuron types.¹⁷ The value of this local complexity likely goes far beyond the current recurrent ANN goals of maintaining a local state for some period of time. In addition to the richness of local biological complexity, there is the consideration of what spike based information processing means with regard to contributing information about time. While there is significant

discussion around spike-based neural algorithms from the perspective of energy efficiency; less frequently noted is the ability of spiking neurons to incorporate information in the time domain. Neuroscience researchers are very familiar with aspects of neural processing for which “when” a spike occurs can be as important as whether a spike occurs at all, however this form of information representation is uncommon in other domains.

Extracting more computational capabilities from spiking and the local circuit complexity seen in cortex and other regions has the potential to enable temporal neural networks to continue to become more powerful and effective in the coming years. However, it is likely the full potential of temporal neural networks will not be fully realized until they are fully integrated into systems that also include the complexity of regional communication in the brain, such as networks configured to perform both top-down and bottom-up processing simultaneously, such as neural-inspired Bayesian inference networks.

3. Bayesian neural algorithms. Even perhaps more than the time, the most common critique from neuroscientists about the neural plausibility of deep learning networks is the general lack of “top-down” projections within these algorithms. Aside from the optic nerve projection from retina to the LGN area of the thalamus, the classic visual processing circuit of the brain includes as much, and often more, top-down connectivity between regions (for example, V2→V1) as it contains bottom-up (V1→V2).

Not surprisingly, the observation that higher-level information can influence how lower-level regions process information has strong ties to well-established motifs of data processing based around Bayesian inference. Loosely speaking, these models allow data to be interpreted not simply by low-level information assembling into higher features unidirectionally, but also by what is expected—either acutely based on context or historically based on past experiences. In effect, these high-level “priors” can bias low-level processing toward more accurate interpretations of what the input means in a broader sense.

While the extent to which the brain is perfectly explained by this Bayesian perspective is continually debated, it is quite clear the brain does use higher-level information, whether from memory, context, or across sensory modalities, to guide perception of any sensory modality. If you expect to see a cloud shaped like a dog, you are more likely to see one. The application of these concepts to machine learning has been more limited, however. There are cases of non-neural computer vision algorithms based on Bayesian inference principles,²² though it has been challenging to develop such models that can be trained as easily as deep learning networks. Alternatively, other algorithms, such as Recursive Cortical Networks (RCNs),¹³ Hierarchical Temporal Memory (HTM),² and predictive networks (PredNet)²⁴ have been developed that also leverage these top-down inputs to drive network function. These approaches are not necessarily explicitly Bayesian in all aspects, but do indicate that advances in this area are occurring.

Ultimately, however, this area will be enabled by increased knowledge about how different brain areas interact with one another. This has long been a challenge to neuroscientists, as most experimental physiology work was relatively local and anatomical tracing of connectivity has historically been sparse. This is changing as more sophisticated physiology and connectomics techniques are developed. For example, the recently proposed technique to “bar-code” neurons uniquely could enable the acquisition of more complete, global graphs of the brain.²⁰

Of course, the concept of Bayesian information processing of sensory inputs, like the previous two algorithmic frameworks described previously, is skewed heavily toward conventional machine learning tasks like classification. However, as our knowledge of the brain becomes more extensive, we can begin to take algorithmic inspiration from beyond just sensory systems. Most notable will be dynamics and memory.

4. Dynamical memory and control algorithms. Biological neural circuits have both greater temporal and architectural complexity than classic ANNs. Beyond just being based on

spikes and having feedback, it is important to consider that biological neurons are not easily modeled as discrete objects like transistors, rather they are fully dynamical systems exhibiting complex behavior over many state variables. While considering biological neural circuits as complex assemblies of many dynamical neurons whose interactions themselves exhibit complex dynamics seems intractable as an inspiration for computing, it is worth noting that there is increasing evidence that it is possible to extract computational primitives from such neural frameworks, particularly when anatomy constraints are considered. Increasingly, algorithms like liquid state machines (LSMs)²⁵ have been introduced that abstractly emulate cortical dynamics loosely by balancing activity in neural circuits that exhibit chaotic (or near chaotic) activity. Alternatively, by appreciating neural circuits as programmable dynamical systems, approaches like the neural engineering framework (NEF) have shown that complex dynamical algorithms can be programmed to perform complex functions.¹⁰

While these algorithms have shown that dynamics can have a place in neural computation, the real impact from the brain has yet to be appreciated. Neuroscientists increasingly see regions like the motor cortex, cerebellum, and hippocampus as being fundamentally dynamical in nature: it is less important what any particular neuron’s average firing rate is, and more important what the trajectory of the population’s activity is.

The hippocampus makes a particularly interesting case to consider here. Early models of the hippocampus were similar to Hopfield networks—memories were represented as auto-associative attractors that could reconstruct memories from a partial input. These ideas were consistent with early place cell studies, wherein hippocampal neurons would fire in specific locations and nowhere else. While a simple idea to describe, it is notable how for roughly forty years this idea has failed to inspire any computational capabilities. However, it is increasingly appreciated that the hippocampus is best considered from a dynamical view: place cell

behavior has long been known to be temporally modulated and increasing characterization of “time cells” is indicative that a more dynamical view of hippocampal memory is likely a better description of hippocampal function and potentially more amenable to inspiring new algorithms.

Of course, developing neural-inspired dynamical memory and control algorithms has the potential to greatly advance these existing techniques, but the real long-lasting benefit from neural computing will likely arise when neuroscience provides the capability to achieve higher-level cognition in algorithms.

5. The unknown future: Cognitive inference algorithms, self-organizing algorithms and beyond. Not coincidentally, the description of these algorithms has been progressing from the back of the brain toward the front, with an initial emphasis on early sensory cortices and eventually progressing to higher level regions like motor cortex and the hippocampus. While neural machine learning is taking this back-to-front trajectory, most of these areas have all received reasonably strong levels of neuroscience attention historically—the hippocampus arguably is as well studied as any cortical region. The “front” of the brain, in contrast, has continually been a significant challenge to neuroscientists. Areas such as the prefrontal cortex and its affiliated subcortical structures like the striatum have remained a significant challenge from a systems neuroscience level, in large part due to their distance from the sensory periphery. As a result, behavioral studies of cognitive functions such as decision making are typically highly controlled to eliminate any early cortical considerations. Much of what we know from these regions originates from clinical neuroscience studies, particularly with insights from patients with localized lesions and neurological disorders, such as Huntington’s and Parkinson’s diseases.

As a result, it is difficult to envision what algorithms inspired by prefrontal cortex will look like. One potential direction are recent algorithms based on deep reinforcement learning, such as AlphaGo’s deep Q-learning,

which has been successful in winning against humans in games of presumably complex decision making.²⁹ These deep reinforcement algorithms today are very conventional, only touching loosely on the complexity of neuromodulatory systems such as dopamine that are involved in reinforcement learning, yet they are already disrupting many of the long-held challenges in artificial intelligence. Still, the reliance of modern reinforcement learning techniques on training methods from deep learning limits their utility in domains where training data is sparse; whereas biological neural circuits can rapidly learn new tasks on comparably very few trials.

While learning more about the frontal and subcortical parts of the brain offer the potential to achieve dramatic capabilities associated with human cognition, there is an additional benefit that we may achieve from considering the longer time-scales of neural function, particularly around learning. Most current neural algorithms learn through processes based on synaptic plasticity. Instead, we should consider that the brain learns at pretty much all relevant time-scales – over years even in the case of hippocampal neurogenesis and developmental plasticity. Truly understanding how this plasticity relates to computation is critical for fully realizing the true potential of neural computation.

We and others have begun to scrape the surface of what lifelong learning in algorithms could offer with early signs of success. For instance, adding neurogenesis to deep learning enables these algorithms to function even if the underlying data statistics change dramatically over time,⁹ although the relative simplicity of deep learning is not suitable to leverage much of the potential seen in the biological process. The recently announced Lifelong Learning in Machines program by DARPA promises to explore this area deeper as well. Ultimately, the implications of these differing learning mechanisms in the brain are highly dependent on the diversity of neural circuit architectures; thus, as algorithms begin to better leverage neural circuits in their design, the opportunities to

extend algorithm function through learning will likely appear.

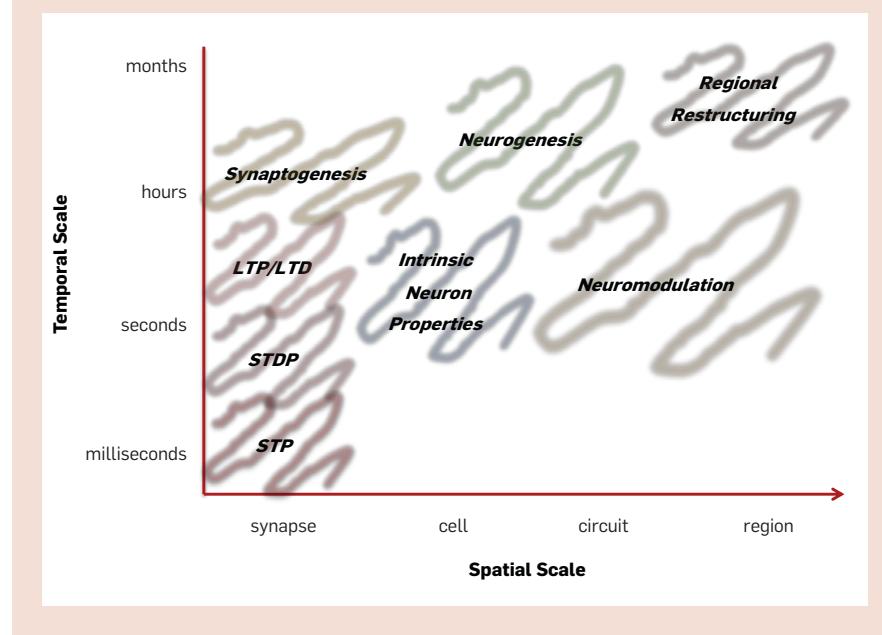
Progress in neural algorithms will build on itself. This progression of neural functionality very well could make combined systems considerably more powerful than the individual components. The neural Turing machine concept took inspiration from the brain's higher-level working memory capabilities by combining neural networks with conventional computing memory resources.¹⁵ It is likely this process can be performed entirely within a neural context—if a hippocampal-inspired one-shot learning algorithm could enable the continuous adaptation of a deep learning network, it could considerably increase the long-term utility of that algorithm. This consideration of amortized cost of an algorithm is of course a very different approach to evaluating the costs and benefits of computing, but the prospect of continuously learning neural systems will require some sort of long-term evaluation.⁹ Further, just as the brain uses the hippocampus to provide a short-term

memory function to complement the long-term memory of several sensory cortices, it is likely that future neural systems could be constructed in a modular manner whereby different combinations of neural components can amplify the performance on different functions.

While this discussion has focused primarily on the long-term benefits of modular neural algorithms, this predicted succession of algorithmic capabilities would be well positioned to be amplified by corresponding advances in computing architectures and materials.¹⁶ Today, deep learning has already begun to substantially influence the design of computer architectures such as GPUs and specialized deep learning architectures such as Google's TPU¹⁸ and implicitly underlying devices and materials. While materials have often been researched with respect to the ubiquitous binary transistor function common to von Neumann architectures, new architectures inspired by novel neural algorithm classes may introduce entirely new desirable characteristics for devices. For example,

Figure 3. Biological neural circuits exhibit learning at many spatial and temporal scales.

At synaptic scales, short-term plasticity (STP) changes synapse strengths at very rapid (spike-to-spike) intervals, whereas spike-timing dependent plasticity (STDP) and long-term potentiation / depression (LTP/LTD) affect synaptic strengths over longer time scales, and are more analogous to neural network learning. Learning is not restricted to existing synapses, with neurons also changing their dynamics in response to inputs, and at longer timescales adding new synapses (synaptogenesis) and neurons (neurogenesis) does occur in select brain regions. Finally, learning occurs at macroscopic scales as well; with neuromodulators affecting neuronal dynamics over large brain regions; and potentially even restructuring of brain regions seen at long timescales in response to injury.



dynamical neural algorithms inspired by prefrontal and motor cortex may be best implemented on more dynamics-friendly devices capable of smooth state transitions as opposed to the very stiff and reliable operational characteristics of transistors today. One particular area where neural architectures could begin to have dramatic impact would be intrinsic capabilities for learning and self-organization. While we are still a long-way away from understanding neural development from a computational theory perspective, the availability of such functionality at an architectural level will likely be very disruptive, particularly as algorithms leveraging more brain-like plasticity mechanisms are introduced.

Can Neuroscience Really Drive Computing Long-Term?

While an argument has been made for why neural computing could provide the computing industry with a future beyond Moore's Law, by no means is this future assured. Aside from the clear technical challenges that lie ahead related to implementing the intellectual trajectory laid out here; there are considerable social challenges that must be addressed as well.

Arguably, the greatest urgency is to inspire the broader neuroscience community to pursue developing theories that can impact neural computing. While there is considerable reason to believe that our knowledge of the brain will continue to accelerate through improved neurotechnologies, the path by which that knowledge can be leveraged into a real impact on computing is not well established. In particular, it is notable that much of the deep learning revolution was driven by computer scientists and cognitive scientists basing algorithms primarily on concepts well established in neuroscience in the 1940s and 1950s. There are several examples to have optimism, however. The IARPA MICrONS program, which is part of the U.S. BRAIN Initiative, aims directly at the challenge of leveraging high-throughput neuroscience data in novel algorithm development.⁶ Google's DeepMind—a company started by cognitive neuroscientists—is at the forefront of successfully integrating neural con-

The greatest urgency is to inspire the broader neuroscience community to pursue developing theories that can impact neural computing.

cepts such as reinforcement learning into machine learning algorithms.²⁹ The EU Human Brain Project has been successful at renewing interest in neuromorphic technologies in the computer science and electrical engineering communities.

Nevertheless, there must be a more robust investment by neuroscientists if computing is to benefit from the revolutions underway in experimental neuroscience. This is particularly important if neural influence is to move beyond computer vision—a community that has long had ties to neuroscience vision researchers. For example, despite a historic level of attention and understanding that is roughly comparable to that of visual cortex,²⁶ the hippocampus has had arguably very little influence on computing technologies, with only limited exploration of hippocampal-inspired spatial processing in simultaneous localization and mapping (SLAM) applications²⁸ and almost no influence on computer memory research.

A renewed focus by neuroscientists on bringing true brain-inspiration to computation would be consistent with the field's broader goals in addressing the considerable mental health and neurological disorders facing society today.⁴ Many of the clinical conditions that drive neuroscience research today can be viewed as impairments in the brain's internal computations, and it is not unreasonable to argue that taking a computing-centric perspective to understanding neurologically critical brain regions such as the striatum and hippocampus could facilitate new perspectives for more clinically focused research.

A second, related challenge is the willingness of the computing communities to incorporate inspiration from a new source. Computing advances have been driven by materials for decades, with reduced emphasis on addressing the underlying von Neumann architecture. Given the perceived plateauing of this classic path, there is now considerable investment in neural architectures; efforts such as IBM TrueNorth²⁷ and the SpiNNaker²¹ and BrainScales³² systems out of the EU HBP have focused on powerful architectural alternatives in anticipation of neural algorithms. Other more-device

driven efforts are focused on using technologies such as memristors to emulate synapses. To some extent, these approaches are seeking to create general purpose neural systems in anticipation of eventual algorithm use; but these approaches have had mixed receptions due to their lack of clear applications and the current success of GPUs and analytics-specific accelerators like the TPU. It is reasonable to expect that new generations of neural algorithms can drive neuromorphic architectures going forward, but the parallel development of new strategies for neural algorithms with new architecture paradigms is a continual challenge. Similarly, the acceptance of more modern neuroscience concepts by the broader machine learning community will likely only occur when brain-derived approaches demonstrate an advantage that appeared insurmountable using conventional approaches (perhaps once the implications of Moore's Law end its reach that community); however, once such an opportunity is realized, the deep learning community is well-positioned to take advantage of it.

One implication of the general disconnect between these very different fields is that few researchers are sufficiently well versed across all of these critical disciplines to avoid the sometimes-detrimental misinterpretation of knowledge and uncertainty from one field to another. Questions such as "Are spikes necessary?" have quite different meanings to a theoretical neuroscientist and a deep learning developer. Similarly, few neuroscientists consider the energy implications of complex ionic Hodgkin-Huxley dynamics of action potentials, however many neuromorphic computing studies have leveraged them in their pursuit of energy efficient computing. Ultimately, these mismatches demand that new strategies for bringing 21st century neuroscience expertise into computing be explored. New generations of scientists trained in interdisciplinary programs such as machine learning and computational neuroscience may offer a long-term solution; but in the interim, it is critical that researchers on all sides are open to the considerable progress made these complex, well-established domains in which they are not trained.

Acknowledgments

The author thanks Kris Carlson, Erik Debenedictis, Felix Wang, and Cooke Santamaria for critical comments and discussions regarding the manuscript. The authors acknowledge financial support from the DOE Advanced Simulation and Computing program and Sandia National Laboratories' Laboratory Directed Research and Development Program. Sandia National Laboratories is a multiprogram laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

This article describes objective technical results and analysis. Any subjective views or opinions that might be expressed do not necessarily represent the views of the U.S. Department of Energy or the U.S. Government. □

References

- Agarwal, S. et al. Energy scaling advantages of resistive memory crossbar based computation and its application to sparse coding. *Frontiers in Neuroscience* 9.
- Ahmad, S. and Hawkins, J. Properties of sparse distributed representations and their application to hierarchical temporal memory. arXiv:1503.07469.
- Association, S.I. and Corporation, S.R. Rebooting the IT Revolution: A Call to Action, 2015.
- Bargmann, C. et al. BRAIN 2025: A scientific vision. *Brain Research Through Advancing Innovative Neurotechnologies Working Group Report to the Advisory Committee to the Director, NIH*. U.S. National Institutes of Health, 2014; <http://www.nih.gov/science/brain/2025/>.
- Bouchard, K.E. et al. High-performance computing in neuroscience for data-driven discovery, integration, and dissemination. *Neuron* 92, 3, 628–631.
- Cepelewicz, J. The U.S. Government launches a \$100-million 'Apollo project of the brain.' *Scientific American*.
- Churchland, A.K. and Abbott, L. Conceptual and technical advances define a key moment for theoretical neuroscience. *Nature Neuroscience* 19, 3, 348–349.
- Dennard, R.H., Gaenslen, F.H., Rideout, V.L., Bassous, E. and LeBlanc, A.R. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid-State Circuit* 9, 5, 256–268.
- Draelos, T.J. et al. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *Proceedings of the 2017 International Joint Conference on Neural Networks*. IEEE, 526–533.
- Eliasmith, C. and Anderson, C.H. *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, Cambridge, MA, 2004.
- Esmailzadeh, H., Blehm, E., Amant, R.S., Sankaranigam, K. and Burger, D. Dark silicon and the end of multicore scaling. In *Proceedings of the 2011 38th Annual International Symposium on Computer Architecture*. IEEE, 365–376.
- Gao, P. and Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology* 32, 148–155.
- George, D. et al. A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science* 358, 6368, eaag2612.
- Goodfellow, I. et al. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014, 2672–2680.
- Graves, A., Wayne, G. and Danihelka, I. Neural Turing machines; arXiv:1410.5401.
- Indiveri, G., Linares-Barranco, B., Hamilton, T.J., van Schaik, A., Etienne-Cummings, R., Delbrück, T., Liu, S.-C., Dudek, P., Häfliger, P. and Renaud, S. Neuromorphic Silicon Neuron Circuits. *Frontiers in Neuroscience*, 5, 73.
- Jiang, X. et al. Principles of connectivity among morphologically defined cell types in adult neocortex. *Science* 350, 6264, aac9462.
- Jouppi, N.P. et al. Datacenter performance analysis of a tensor-processing unit. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*. ACM, 2017, 1–12.
- Kasthuri, N. et al. Saturated reconstruction of a volume of neocortex. *Cell* 162, 3, 648–661.
- Kehschnull, J.M., da Silva, P.G., Reid, A.P., Peikon, I.D., Albeau, D.F. and Zador, A.M. High-throughput mapping of single-neuron projections by sequencing of barcoded RNA. *Neuron* 91, 5, 975–987.
- Khan, M.M. et al. SpiNNaker: Mapping neural networks onto a massively-parallel chip multiprocessor. In *Proceedings of the IEEE 2008 International Joint Conference on Neural Networks*. IEEE, 2849–2856.
- Lake, B.M., Salakhutdinov, R. and Tenenbaum, J.B. Human-level concept learning through probabilistic program induction. *Science* 350, 6266, 1332–1338.
- LeCun, Y., Bengio, Y. and Hinton, G. Deep learning. *Nature* 521, 7553, 436–444.
- Lotter, W., Kreiman, G. and Cox, D. Deep predictive coding networks for video prediction and unsupervised learning. arXiv:1605.08104.
- Maass, W., Natschläger, T. and Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation* 14, 11, 2531–2560.
- Marr, D. Simple memory: A theory for archicortex. *Philosophical Trans. Royal Society of London. Series B, Biological Sciences*, 23–81.
- Merolla, P.A. et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science* 345, 6197, 668–673.
- Milford, M.J., Wyeth, G.F. and Prasser, D. RatSLAM: A hippocampal model for simultaneous localization and mapping. In *Proceedings of the 2004 IEEE International Conference on Robotics and Automation*. IEEE, 403–408.
- Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* 518, 7540, 529–533.
- Moore, G.E., Progress in digital integrated electronics. *Electron Devices Meeting*, (1975), 11–13.
- Nikonov, D.E. and Young, I.A. Overview of beyond-CMOS devices and a uniform methodology for their benchmarking. In *Proceedings of the IEEE* 101, 12, 2498–2533.
- Schemmel, J., Briiderle, D., Griebl, A., Hock, M., Meier, K. and Millner, S. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*. IEEE, 1947–1950.
- Shalf, J.M. and Leland, R. Computing beyond Moore's Law. *Computer* 48, 12, 14–23.
- Shor, P.W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Review* 41, 2, 303–332.
- Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Machine Learning Research* 15, 1, 1929–1958.
- Stevenson, I.H. and Kording, K.P. How advances in neural recording affect data analysis. *Nature Neuroscience* 14, 2, 139–142.
- Thompson, S.E. and Parthasarathy, S. Moore's Law: The future of Si microelectronics. *Materials Today* 9, 6, 20–25.
- Waldrop, M.M. The chips are down for Moore's Law. *Nature News* 530, 7589, 144.
- Williams, R.S. and DeBenedictis, E.P. OSTP Nanotechnology-Inspired Grand Challenge: Sensible Machines (ext. ver. 2.5).
- Yuste, R. From the neuron doctrine to neural networks. *Nature Reviews Neuroscience* 16, 8, 487–497.

James B. Aimone (jbaimon@sandia.gov) is Principal Member of Technical Staff at the Center for Computing Research, Sandia National Laboratories, Albuquerque, NM, USA.

Copyright held by author/owner.

review articles

DOI:10.1145/3241037

Quantum systems will significantly affect the field of cyber security research.

BY PETROS WALDEN AND ELHAM KASHEFI

Cyber Security in the Quantum Era

CYBER SECURITY DEALS with the protection of computer systems from attacks that could compromise the hardware, software or information. These attacks, by allowing unauthorized use, could leak private information and cause damage or disruption. In the future, the part of everyday life and economy requiring computer systems is bound to increase further and become fully dominant. Cyber warfare and cyber crime will be common and the role of cyber security crucial.

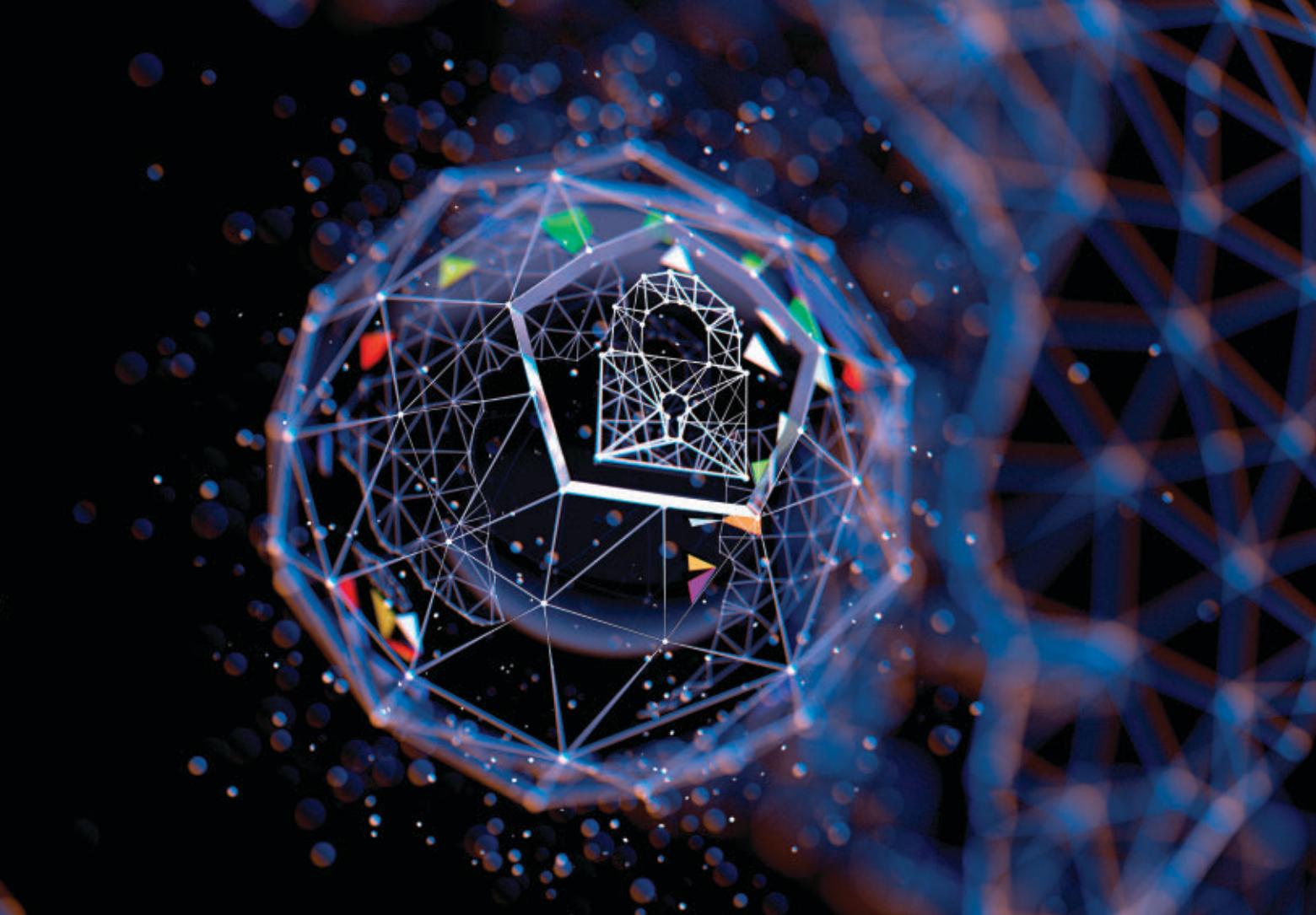
SINCE THE COMPUTER systems (and attackers) evolve both in hardware and software, the constant evolution of this field is of high importance. Arguably the most dramatic development that one can envision is a change in the paradigm of computational model used. Quantum technologies appear to bring us close to such a change. Here, we explore the research field that lies on the intersection of cyber security and quantum technologies research.

The dawn of the quantum technologies era. One of the major scientific revolutions of the 20th century was the development of *quantum theory*. From its early days,³³ all the way until the development of the full mathematical formalism³⁸ and the subsequent development of first wave of applications (for example, transistors, laser, superconductors, among others) quantum theory has been very successful in many different settings being confirmed in unprecedented accuracy (record accuracy of 10^{-8} for the anomalous magnetic dipole moment).²⁴ Crucial in this first wave of applications was the new understanding of nature that quantum theory provided. However, the ability to control quantum systems as desired was limited, putting restrictions on the class of technological applications that one could envision.

In recent years this has changed¹⁶ and the control of quantum systems has advanced considerably, while further progress appears very plausible in the near future due to the increased interest and investments as well as the scientific breakthroughs that have already occurred. Many countries all around the globe have launched national

» key insights

- Quantum computers will pose a significant threat for cyber security. When large fault-tolerant quantum computers are constructed the most commonly used cryptosystems will break. Therefore dealing with this threat is crucial and timely.
- Securing fully classical protocols against quantum-technology-equipped adversaries is possible but requires extra care that goes beyond a careful choice of cryptosystems.
- Quantum technologies will also have a positive impact on cyber security. Quantum devices with current state-of-the-art technology can be used to enhance the security by achieving tasks impossible classically, such as, secret-key expansion with perfect security. Since quantum computers will become an integral part of our future network of communications and computations, we need to develop practical ways to use the quantum computers with same security guarantees with those of secure (classical) computing.



quantum technologies programs, varying from millions to billions, including those of Australia, Canada, China, EU, Japan, Netherlands, Russia, Singapore, U.K., U.S. At the same time, major industrial players such as Google, IBM, Microsoft, Intel, Atos, Baidu, Alibaba, Tencent along with numerous smaller and bigger quantum start-ups have initiated labs developing quantum hardware and software. This has led to what is now called “the second quantum revolution,” where the ability to manipulate quantum systems as desired is leading to an era in which a variety of new technologies will appear and, in certain cases, could potentially replace existing solutions.

Arguably, the most important quantum technology will be the development of computation devices that exploit quantum phenomena, which we refer to as *quantum computers*. Quantum computers are likely to become a disruptive innovation as they can offer considerably greater computational power than their classical counterparts.

Here, we must stress that this is not

something that will become relevant in the far future. Impressive quantum technological achievements are already available. To name two recent examples: Google’s latest quantum processor “Bristlecone” has a record of 72 qubits with very low error rates, and is expected to be larger in size than what the best classical supercomputers can simulate.³ Satellite quantum key distribution has been realized, enabling information theoretic secure encryption over distances of 7600km (intercontinental) and used as basis for a secure teleconference between the Austrian Academy of Sciences and the Chinese Academy of Sciences.²⁷

Quantum cyber security. The development of large quantum computers, along with the extra computational power it will bring, could have dire consequences for cyber security. For example, it is known that important problems such as factoring and the discrete log, problems whose presumed hardness ensures the security of many widely used protocols (for example, RSA, DSA, ECDSA), can be solved ef-

ficiently (and the cryptosystems broken), if a quantum computer that is sufficiently large, “fault tolerant” and universal, is developed.³⁵ While this theoretical result has been known since the 1990s, the actual prospect of building such a device has only recently become realistic (in medium term). However, addressing the eminent risk that adversaries equipped with quantum technologies pose is not the only issue in cyber security where quantum technologies are bound to play a role.

Quantum cyber security is the field that studies all aspects affecting the security and privacy of communications and computations caused by the development of quantum technologies.

Quantum technologies may have a negative effect to cyber security, when viewed as a resource for adversaries, but can also have a positive effect, when honest parties use these technologies to their advantage. The research can, broadly speaking, be divided into three categories that depend on who has access to quantum technologies and how developed these technologies

are (see Figure 1). In the first category we ensure that currently possible tasks remain secure, while in the other two categories we explore the new possibilities that quantum technologies bring.

As is typical in cryptography, we first assume the worst-case scenario in terms of resources, where the honest parties are fully classical (no quantum abilities), while the adversaries have access to any quantum technology (whether this technology exists currently or not). In particular we assume they have a large quantum computer. Ensuring the security and privacy guarantees of a classical protocol remain intact is known as *post-quantum* (or “quantum-safe”) security.

In the second category we allow honest parties to have access to quantum technologies in order to achieve

enhanced properties, but we restrict this access to those quantum technologies that are currently available (or that can be built in near-term). Requesting this level of quantum abilities comes from the practical demand to be able to construct now, small quantum devices/gadgets that implement the “quantum” steps of (the honest) protocols. The adversaries, again, can use any quantum technology. In this category we focus on achieving classical functionalities but we are able to enhance the security or efficiency of the protocols beyond what is possible classically by using current state-of-the-art quantum gadgets.

Finally, the third category looks further in the future and examines the security and privacy of protocols that are possible (are *enabled*) by the existence

of quantum computers. We assume there exist quantum computation devices that offer advantages in many useful applications compared with the best classical computers. At that time, there will be tasks that involve quantum computers and communication and processing of quantum information, where the parties involved want to maintain the privacy of their data and have guarantees on the security of the tasks achieved. This period may not be too far, since quantum devices being developed now are already crossing the limit of quantum computations that can be simulated by classical supercomputers.

These categories, in general, include all aspects of cyber security. We will focus on the effects that quantum technologies have for cryptographic attacks and attacks that exploit vulnerabilities of the new quantum hardware when such hardware is used. As far as exploits of other vulnerabilities of existing classical hardware is concerned (for example, timing attacks), we do not expect they will significantly benefit from quantum technologies and thus we do not expand further.^a

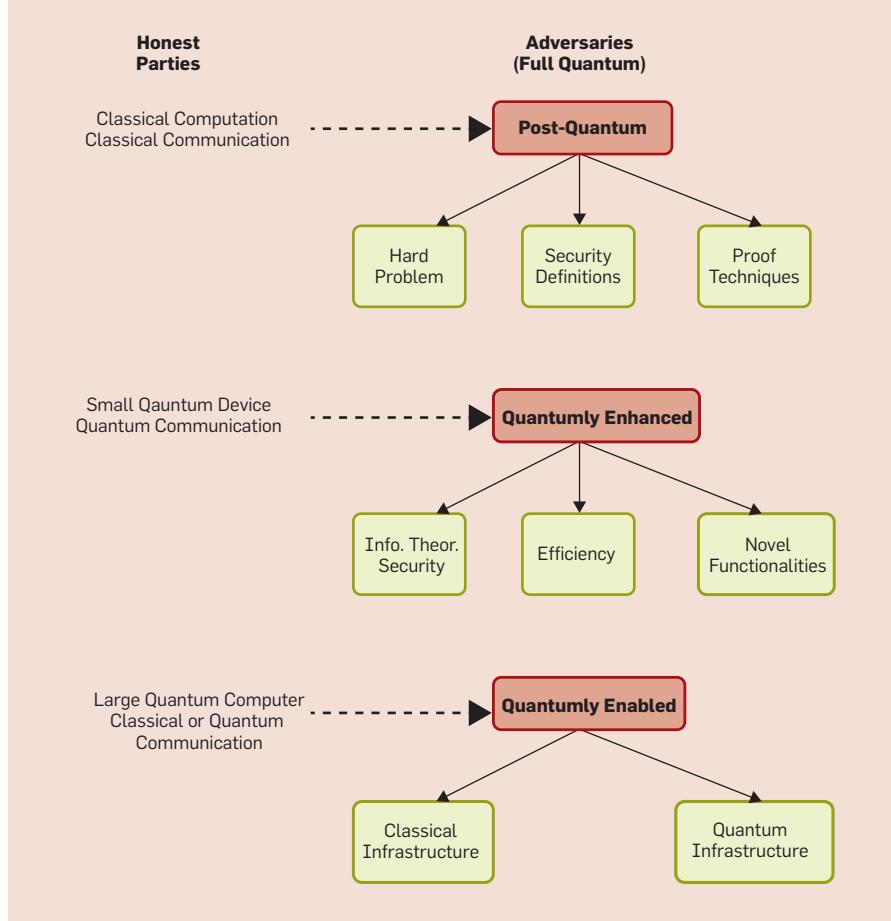
This review. First of all we clarify what this review is not. It is not an exhaustive list of all research in quantum cyber security, neither a historical exposition on how quantum cryptography developed, nor a proper introduction to the field including the background required. Excellent such reviews have been written (for example, Broadbent¹³).

Our aim in this article is twofold. On the one hand, we want to clarify misconceptions and organize/categorize the research landscape in quantum cyber security in a comprehensible and approachable way to the non-experts. On the other hand, we want to focus on specific aspects, for each of the quantum cyber security research categories given here that we believe have been underrepresented in research and exposure to the public, despite being very important. We clarify some facts about quantum computers and quantum adversaries, setting the stage to ana-

^a One could imagine that enhanced quantum sensing and quantum metrology could improve certain side-channel attacks, but this is beyond the scope of this article.

Figure 1. Schematic representation of the quantum cyber security research landscape.

Red boxes are the three categories of research. Dotted lines indicate the resources (computation and communication) required from the honest parties. Green boxes represent issues on which we focus in this review. For the post-quantum category, we consider the changes required: which are the hard problems used, security definitions and proof techniques. For the quantumly enhanced category we consider the types of enhancements we may get in different protocols: information theoretic security (from computational), increased efficiency, functionalities impossible classically (even with computational assumptions). For the quantumly enabled category we consider separately the different communication infrastructures available (classical/quantum).



lyze the three categories of quantum cyber security research. We explore post-quantum security, giving a brief overview of the field and focusing on the issue of security definitions and proof techniques. Then we sketch the research directions in quantum-enhanced security, focusing on the issue of implementation attacks and device independence. Later, we focus on classical clients securely delegating computations to the quantum cloud and conclude with a glimpse of how we envision cyber security will be reshaped in the decades to come.

Myths and Realities about Quantum Computing

In many popular accounts, quantum computers are described as some mythical computation devices that, if ever constructed, would magically solve pretty much anything one can imagine in a fraction of a second. In reality, the power of quantum computers is much more modest. Here, we clarify four of the most common misconceptions on the computational power and possibilities of quantum computers and quantum adversaries. In this way we can see the effects on cyber security research more clearly.

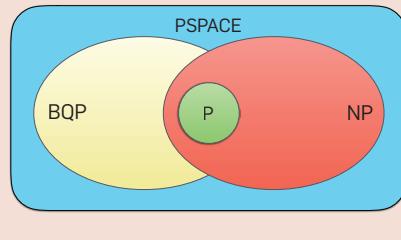
Myth 1. *Quantum computers are much faster in performing operations than classical computers.*

Reality. Quantum computers are not faster in the sense of implementing larger number of operations per second. The computational speed-up that quantum computers offer is achieved because quantum theory allows algorithms that include operations practically impossible for classical computers. Therefore, to achieve a speed-up is a task that requires the invention of new algorithms that use these operations suitably and is not straightforward. Indeed, the exact speed-up highly depends on the specific problem considered. This problem-dependency of the speed-up justifies why a quantum adversary can break only certain public-key cryptosystems, while others may remain secure with minor modifications (for example, in the key lengths).

Myth 2. *Quantum computers simultaneously perform all branches of a (probabilistic) computation and can find accepting paths instantly.*

Figure 2. Conjectured relation of complexity classes.

The conjectured relations are reinforced by the existence of oracle results that separate BQP with NP.



Reality. Quantum computers span the space of possibilities, computational branches, in a peculiar way. It is similar with classical probabilistic computers (BPP),^b with the important difference that quantum computers behave as having “probabilities” that take complex values. This behavior, leads to “cancellations” of certain branches, since adding complex numbers is not monotonically increasing (unlike adding numbers in the interval [0, 1] as for BPP devices). This property, along with algorithms that exploit it, leads to quantum speed-ups. However, at the end of a quantum computation, the result (accepting/rejecting in a decision problem) is obtained by a single read-out/measurement and, therefore, all the “unrealized” branches do not contribute, contrary to the myth that quantum computers perform all branches in parallel in a way that someone can meaningfully extract all the information present in those branches.

Myth 3. *Quantum computers can efficiently solve NP-complete problems (such as Traveling Salesman Problem).*

Reality. The class of decision problems that quantum computers can solve efficiently is called BQP. Its (conjectured) relation to other known classes can be seen in Figure 2. In particular, we see that NP is *not* contained in BQP and, therefore, a quantum computer cannot solve a NP-complete problem efficiently. Having said that, it is important to note that quantum computers may (and do) provide polynomial or

constant speed-up in many problems outside BQP (for example, quadratic search speed-up), including such speed-ups for NP-complete problems. For many tasks, even such a moderate speed-up could be of great importance. In cyber security, for example, it affects the size of keys needed to guarantee a requested level of security.

Myth 4. *Using problems that are hard for a quantum computer (outside BQP) suffices to make a cryptographic protocol secure against any quantum attack.*

Reality. This is a necessary but not sufficient condition. A quantum attacker can use quantumness in various ways, not only in order to solve some classical problems quicker. Next, we give such examples (superposition attacks) and specify that both the security definitions and the proof techniques need to be modified.

Post-Quantum Security: Quantum Adversaries

Even if one is not convinced of the necessity to drastically change the existing cryptographic protocols and infrastructure to use quantum technologies in a positive way, they still must address the possibility that current or future adversaries might use such technologies to their benefit. Since scalable, fault-tolerant, quantum computers that could break current cryptography require thousands of qubits, one could assume this is unlikely to happen in this decade.

However, there are three important reasons why we need to address quantum attackers now. Firstly, security can be broken retrospectively. For instance, if some agency intercepts and stores encrypted email messages sent today, and 10 years later develops a quantum computer, they can then use it to decrypt them.¹ Secondly, to develop cryptographic solutions that are post-quantum secure, to achieve high efficiency and to build confidence on the security of these solutions is an endeavor that requires years of research by multiple, independent, top research groups. Thirdly, changing the cryptographic infrastructure will also require years once we have decided to do it.

We will divide the research in post-quantum cryptography in three classes, according to the ways we allow the adversaries to use their “quantum

b It is the class of decision problems that can be solved efficiently by a probabilistic Turing machine with error bounded away from 1/2. This class captures well the problems that can be solved efficiently by modern classical computers.

abilities" (see Figure 1). The first class is where the adversaries are classical with the extra ability to also solve problems in BQP. In other words, these adversaries are like the standard classical adversaries, with extra access to an oracle/quantum-computer. This class is the best known and, in many accounts of post-quantum security, the only one discussed. In the second and third classes we give extra abilities to the adversaries, for example, we allow them to send input (queries) in quantum states (superposition of classical bits) and then use the (quantum) output, along with their quantum computer oracle, to compromise the security of protocols. The second class addresses the modeling and modification of security definitions, along with the immediate consequences. In the third class we deal with the changes required in (involved) proof techniques in this new quantum security model. For example, the internal state of an adversary during an interactive protocol is described by a general quantum state and their actions by a general quantum map.

We should note that all protocols in the post-quantum security category are, in terms of technologies used, fully classical protocols. All the honest steps are realized in classical devices. Still, understanding quantum computation (hardness of problems) and quantum technologies in general (modeling other types of attacks) is crucial to prove the *security*.

Security against an adversary with an oracle quantum computer. The first, and by far the most thoroughly researched area, is to ensure the security of protocols used is based on the hardness of problems which remain hard for quantum computers. This is clearly a priority since, once quantum computers are built, attacking cryptosystems that use problems that are not hard for quantum computers will be trivial. As explained in Myth 3 and depicted in Figure 2, it is believed that there exist NP problems outside BQP, therefore public key cryptography is still possible in the setting that adversaries have access to an oracle quantum computer.

There are many cryptosystems believed to be resistant to attacks of this type. They can be divided to hash-

based, code-based, lattice-based, multivariate, and secret-key cryptography (see details in Bernstein⁷). Here, we will comment on three issues: confidence, usability, and efficiency.

The belief that a problem is hard, while in some cases it comes from theoretical implications that involve containments of complexity classes, is frequently based on the inability to find efficient solutions (or improve on existing solutions) despite the effort of many groups during a long period of time. This is the case for the hardness of factoring for classical computers. When we examine the hardness of problems used in cryptosystems against *quantum* computers, the confidence we have is generally even smaller. For example, with the exception of Merkle's hash-tree public signature system and McEliece's hidden-Goppa-code public-key encryption system, all other post-quantum proposals are relatively new. What is even more important is that research in quantum algorithms and quantum complexity theory is also new and proper cryptanalysis of the systems from the perspective of a quantum adversary is not yet as thorough as for the classical case.

This brings us to the issue of standardization and usability. Such initiatives are active, for example, the National Institute of Standards and Technology (NIST) had a recent call to standardize post-quantum public-key cryptosystems. In order to define the quantum-bit security of a cryptosystem, one must establish which is the fastest algorithm that attempts to break the system. This would also determine what key-length is required for a given security level. For example, the quadratic speed-up due to Grover's algorithm lead to a need for keys of double size. This, of course, would change if a better algorithm is invented, but good candidates for standardized use should be well enough researched to build confidence that (even moderate) speed-ups are not to be found continuously. In contrast, only recently a new speed-up was found for problems using multivariate quadratic equations.¹⁸ Furthermore, after the standardization of encryption functions, one still needs software implementations that are suitable for integration into a variety of applications.

Finally, possibly the greatest challenge, is the issue of efficiency. While for certain applications (defense, financial market, among others) the highest security is desired even with the cost of worse performance, for a great range of everyday applications slowing down the services is not acceptable (people frequently prefer insecure but high-speed services). Existing post-quantum cryptosystems, when taking all aspects into account, lack efficiency (public key-size, signature size, speed of encryption and decryption algorithms, speed of key generation algorithm, and so on). Improving this, or identifying applications that can tolerate one of these aspects being less efficient, is an active field of research.

Superposition attacks: Modifying security notions. Security is frequently defined in terms of the probability that an adversary can succeed in certain hypothetical, interactive games. For example, one defines indistinguishability as a game that adversaries cannot win with probability higher than $1/2$, indicating that randomly guessing the plaintext bit is the best they can do. In this game, the adversary is given extra ability to use a learning phase in which they can request ciphertexts of their chosen plaintexts. The motivation for giving these extra abilities is that one can imagine a scenario that an adversary could persuade an honest party to encrypt a message of their choice. To ensure privacy, this action should not give the adversary any advantage in trying to decrypt other, unknown to the attacker, messages. Now we want to consider adversaries that have the extra ability to make *quantum* queries (and receive quantum answers) in such a game. Mathematically, if the encryption (for example) is described by a function f_k , a "classical" query a quantum party can make is described as $|x\rangle|y\rangle \rightarrow |x\rangle|f_k(x) \oplus y\rangle$, where we note the first register maintains the information on x since quantum (unitary) operations are necessarily reversible. However, a quantum adversary could have initiated the query in superposition $\sum_{x,y} \Psi_{x,y} |x\rangle|y\rangle$ which by quantum linearity would lead to a superposition ciphertext $\sum_{x,y} \Psi_{x,y} |x\rangle|f_k(x) \oplus y\rangle$. The quantum adversary could attempt to use these superposition ciphertexts to break a cryptosystem. It is worth stressing that having a superposition is not the same

as having access to all the terms of the superposition, in the same way that in Myth 2 we explained not all paths are realized. For example, if one measured directly this superposition they would receive a single ciphertext of one (randomly chosen) plaintext and the security would not be compromised. Instead, attacks involve using this output superposition of ciphertexts state in another quantum algorithm that would reveal hidden structures of the cryptosystem.

Requesting from protocols to be secure against this type of attacks leads to new security definitions for a number of functionalities (for example, quantum indistinguishability). Boneh⁸ was the first paper to offer such definitions, where the quantum random oracle model was defined.^c Since then encryption, signatures, pseudo-random functions and message authentication codes have been similarly defined.⁹

Interestingly, there have been attacks of this type to symmetric cryptosystems recently, that provide an exponential speed-up.²⁵ This was the first quantum attack with exponential speed-up to symmetric cryptosystems, using Simon's algorithm. It does not use Shor's or Grover's algorithm^d that are the quantum algorithms typically used to attack cryptosystems, and demonstrates that post-quantum security is much more subtle than generally believed. It is important to require these higher notions of security and to review all candidate post-quantum cryptosystems in the light of quantum cryptanalysis as described.

We should comment on an obvious objection that one could raise, namely that our systems are classical and therefore applying (for example) the encryption algorithm "coherently" on a superposition input, seems not physically motivated.^e We use a hypothetical ex-

In order to define the quantum-bit security of a crypto system, one must establish which is the fastest algorithm that attempts to break the system.

ample of "Frozen Smart-Card" given in Gagliardoni²¹ to demonstrate that one should not ignore this type of attacks even now. Real-world encryption and authentication is frequently implemented on small electronic devices such as smart cards. A quantum attacker could implement a side-channel attack where they get hold of the smart card and attempt to freeze it to a temperature that starts behaving quantum mechanically. Then they attempt to query it in superposition. This type of side-channel attack is not very different to side-channel attacks considered on cryptographic hardware in today's labs, using thermal or electromagnetic manipulation.

Proof techniques against quantum adversaries. To take the most general view, we should model the internal space of a quantum adversary as a generic quantum state and all their actions and communication (with honest parties) as generic quantum operations. Modeling the adversary quantumly has two effects. On the one hand it gives the adversary more ways to deviate/attack, as for example in the superposition attacks described previously. On the other hand, it has an effect on how to prove security since simulating their view requires simulating quantum rather than classical processes. Note that showing that a proof technique is not applicable does not mean finding an attack that breaks the corresponding cryptosystem, it only means it is no longer provably secure.

For example, to define and prove security in functionalities such as secure multiparty computation (SMPC) the concept of simulation is frequently used. In particular, an adversary should be unable to distinguish whether they are interacting with the real (honest) parties or a simulated view that has no direct access to the private information of the parties involved.²⁸ However, since the internal space of the adversary and the interaction with the simulator are quantum the simulated view should also be quantum. This is incompatible with existing constructions of simulators, where the steps taken are not possible for general quantum states.

The key example that we implicitly assume having classical systems when constructing the simulators is the "rewinding" step (for example, see Lindell²⁸ for a detailed explanation of the simula-

^c The (classical) random oracle is an oracle that to each call it responds with a random response. It is used, for example, as a mathematical abstraction to capture the idea of cryptographic hash functions in security proofs.

^d Attacks using Grover's algorithm were performed, for example, on the original Merkle's key exchange, and only recently post-quantum secure modification was developed.

^e Such attacks will be crucial in the future when the infrastructure will, by default, allow for quantum information processing, but here we argue that even before that, these attacks could be implementable.

tion proof technique and the role of rewinding). In rewinding, the simulator is given the ability to copy the internal state, and in some cases to rewind it to an earlier step. However, due to the no-cloning theorem we know that general quantum states *cannot* be copied. This problem was identified early³⁷ while two weaker versions of quantum rewinding, attempting to fix this issue, have been later developed. The first is the oblivious quantum rewinding,³⁹ where one can rewind but is not allowed to “remember” the transcripts of the previous runs apart from whether a rewinding was necessary or not. The second is the special quantum rewinding,³⁶ where by demanding some extra conditions (special and strict soundness) one can retain information from two runs of the rewinding process. Using quantum-rewinding steps comes at a cost, since it can only be proven that the rewound state is close but not exactly the same as the non-rewound state. This leads to a (small) distinguishing possibility between the real and simulated views, affecting the security parameters of the protocol.

These (limited) quantum-rewinding steps can be used to achieve important primitives. In particular the oblivious rewinding was used to prove quantum security of zero-knowledge proofs, while the special quantum rewinding to prove quantum security of proofs of knowledge.

Another application of the quantum rewinding is as subroutine in certain proof techniques. In order to prove security one frequently proves the security against a very weak (essentially honest) adversary and then adds a mechanism that enforces fully malicious adversaries to behave as the weak adversary or else abort. Such techniques are the Goldreich-Micali-Wigderson (GMW) compiler and the cut-and-choose technique. Both can be used to prove the security of Yao’s seminal secure two-party computation protocol against malicious adversaries, and both require rewinding. The GMW compiler uses zero-knowledge proofs and thus the quantum secure version already mentioned suffices, while the cut-and-choose technique can also be proven quantum secure using the special quantum rewinding.²⁶

To recap, considering fully quantum adversaries has consequences in security definitions, proof techniques

We should model the internal space of a quantum adversary as a generic quantum state and all their actions and communication (with honest parties) as generic quantum operations.

and methods used, and on the cryptanalysis of existing protocols beyond what is implied from giving an oracle access to a quantum computer.

Quantumly Enhanced Security: Quantum Gadgets for Classical Parties

Quantum technologies can also offer advantages for cyber security research. To view this positive aspect, we should consider the possibility of including quantum steps in (honest) protocols with the aim of achieving certain improvement compared to the corresponding fully classical setting. The fact that improvements are in principle possible is well established, with the best known example that of quantum key distribution (QKD).⁶ In QKD, using untrusted quantum channels and classical authenticated channels one can establish a shared, secret key between two spatially separated parties, with information theoretic security. This task, essentially information theoretic secure key expansion, is impossible using only classical communications. Importantly, having a protocol with information theoretic security means the security is not based on *any* computational assumption and therefore remains secure even in the presence of an attacker with a quantum computer. Another example of quantum enhancement is the quantum fingerprints,⁴⁰ where two parties can establish if they are sharing the same bit-string using minimal communication. The best classical communication complexity is $\Omega(\sqrt{n})$, which is exponentially more than the quantum communication of $O(\log_2 n)$.

We are interested in using “quantum gadgets,” usually with simple quantum devices (available with current state-of-the-art technologies), to boost classical protocols in a number of ways. The types of enhancement/advantages offered could be, for example, information theoretic security from a computationally secure classical protocol (as in QKD); “computational” security against quantum attackers from no security against quantum or classical attackers;^f

^f For example, in position verification one can achieve a quantum protocol with security against adversaries with bounded amounts of shared entanglement,¹⁴ while no fully secure classical position verification protocol exists, against multiple colluding adversaries.

and improved efficiency, that is, achieving tasks with fewer resources (as in quantum fingerprinting).

The majority of research on quantumly enhanced security is done on QKD, however, many other protocols, functionalities, and primitives exist that admit enhancement and that require similar or slightly more involved quantum technologies. Some of these technologies include: quantum random number generators, quantum fingerprinting, quantum digital signatures, quantum coin flipping, e-voting, Byzantine agreement, quantum money, quantum private information retrieval, secure multiparty computation (SMPC), and position verification.

Since quantum technologies develop rapidly, the possibilities of practical quantum gadgets increase, as more and more quantumly enhanced protocols become realistic. For example, on top of simple quantum communication between parties, we can now have each of the parties having small quantum processors. It is therefore an exciting time for this type of research since we can now consider tailor-made constructions to enhance the performance of specific involved cryptographic protocols such as e-voting or SMPC.

Practicality. Research in this category involves quantum technologies that are currently possible. While this requirement makes such applications possible, for adaptation of quantumly enhanced solutions for wide use, one must establish the necessary infrastructure, namely a reliable and wide quantum communications network. The development of quantum internet is more than a vision for the future, since a big initiative pushing towards this direction is currently under development ("Quantum Internet Alliance").² In the meantime, priority should be given to applications that involve few parties and do not require a fully developed quantum network.

Quantum hacking. The use of quantum gadgets opens the possibility for new attacks, specific to the physical implementations. Standard side-channel attacks (for example, timing) may be less applicable, but there are new side-channel attacks specific to the quantum devices. The best known quantum hacking attacks are the photon number splitting and beam-splitting attacks, both exploiting the

fact that the real systems used for qubits are not single photons as they are modeled theoretically.¹⁰ The thermal blinding of detectors that leak to the adversary information on the measurement choices before the classical post-processing phase,³⁰ something that invalidates the security proof. The latter attacks have been realized against (previous versions) of the commercially available QKD systems of ID Quantique and MagiQ Technologies. Naturally these attacks are specific to each of the implementations of the quantumly enhanced protocols.

Countermeasures discovered to "fix" systems after side-channel attacks come at a cost (for example, better single photon sources or protocols involving decoy states or monitoring the detectors), but other side-channel attacks are likely to appear. Interestingly, quantum theory offers a theoretical method to deal with all side-channel attacks on the quantum gadgets with some extra cost in resources.

Device-independence. What enables side-channel attacks is the mismatch between the ideal modeling of the (quantum) device and the real implementation. One of the most exciting new possibilities that quantum theory offers is that using the fundamental property of quantum non-locality one can achieve quantum cryptographic tasks based only on the classical statistics/correlation of the measurement outcomes, without the need to make any assumption on the (quantum) devices used.¹⁷ In particular, security is maintained even if the devices were prepared by adversaries and given as black-boxes to the honest parties. Device-independent protocols are secure against any side-channel attack on the quantum device and have been developed for many functions: QKD, QRNG, among others. These protocols come with some cost in resources, currently too high for practical use. However, based on weaker correlations, one can make protocols that are secure without trusting some, but not all, devices with considerably reduced cost compared to fully device independent protocols. For example, in measurement-device independent protocols,^{11,29} security is maintained without trusting the measuring/detecting device and thus one avoids the thermal blinding detector

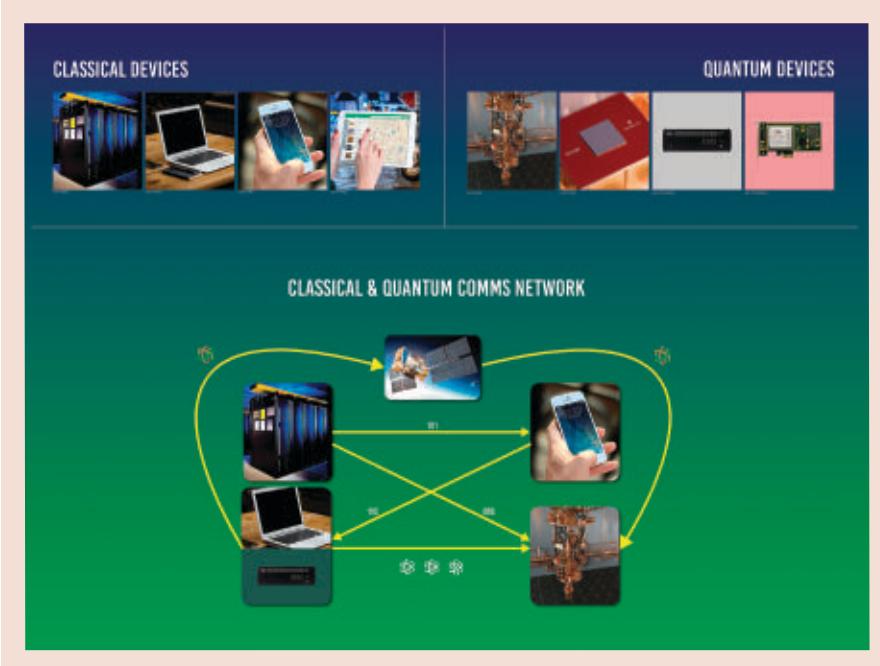
attacks mentioned earlier. In general, there is a trade-off between the extra cost in resources and the amount of trust assumed on quantum devices.

Standardization. For the adoption of quantumly enhanced solutions by industry it is important to establish standards for quantum gadgets compatible and acceptable by the general cyber security community. For QKD, discussions already exist for example within the European Telecommunications Standards Institute (ETSI), while in the case of quantum random number generators, ID Quantique, offers the Quantis product that is validated with the AIS31 methodology. It is important and timely to address the standards issues for all quantumly enhanced functionalities.

Quantumly Enabled Security: Secure Use of Quantum Computers

As we have seen, quantum computers will offer computational advantages in many problems, varying from exponential to much more modest quadratic or even constant. It is natural that when such devices are available, one might want to use this extra computational power in tasks that also require privacy and security, in other words we seek security for quantumly enabled protocols. All security concepts, such as authentication, encryption but also more involved concepts as computation on encrypted data and secure multiparty computation, would need to be modified to apply to *quantum* information and *quantum* computation. Of course, for this type of question to be meaningful, we first must have quantum computation devices of size that can offer concrete computational advantages for everyday problems. This is not the case today, but since we are expected to cross the classical simulation limit (real quantum computers that exceed in size those that can be simulated by classical supercomputers) soon, we are entering the era that will have realistic quantum speed-ups. The time for speed-ups being applied to important everyday problems might not be too far away.

Research in this category is developing rapidly, and already a number of protocols exist for quantum encryption, quantum authentication, quantum non-malleability, blind

Figure 3. The future communication and computation networks.

quantum computation, quantum fully homomorphic encryption, secure multiparty quantum computation, functional quantum encryption, and so on (for example, see the review of Fitzsimons¹⁹). There is a variety of protocols optimizing with respect to different figures of merit, for example, minimizing the quantum (or classical) communication, minimizing the overall quantum resources or the quantum resources of some specific parties, offering the highest possible level of security (information theoretic vs. post-quantum computational).

The majority of these protocols require quantum communication between parties and in most cases, quantum computation must be applied on the communicated quantum information. This raises two concerns: one theoretical and one practical. To achieve such tasks one needs quantum computation devices that are compatible with the quantum communication devices. On the one hand, the best platform for quantum communication is photonic, since it is simple to send quantum information encoded in photons in long distances. On the other hand, one of the most promising approaches for quantum computation devices, the one used by the major industrial players and that is leading the “bigger quantum computer” race, is based on supercon-

ducting qubits. The preferred types of qubits, for communication and computation, do not coincide, and moreover, currently it is not even known if they are compatible. It is unknown if superconducting quantum computers can be part of a “networked architecture,” since they are currently built in a monolithic architecture and is not clear if it will ever be possible to send and receive quantum states.

The practical concern is the two quantum technological developments, namely the quantum computing devices and the large quantum network, are independent and we should be able to use “local” quantum computation devices before establishing the infrastructure required for a full quantum Internet network. For example, even if a single quantum computer is built in some central university or company lab, we may wish to use it to delegate computations before establishing a quantum network infrastructure. This is precisely the case for some of the current, small-scale, quantum computers (IBM, Rigetti); they offer their quantum computer in a cloud service to the public using a classical interface. Therefore, we turn to a question of both practical and theoretical interest: Can we provide quantum computation protocols that *maintain privacy and security guarantees* using this classical interface, that is, to clients with no-quantum abilities, and

what would be the cost of it? This is a question that very recently has attracted attention and following the first key research we analyze here, we expect that it will become very important.

Here, we refer to “blind quantum computation” as all the protocols where a client with no quantum-computing device delegates a computation to a server with a quantum computer maintaining the privacy of her input/output. There is strong evidence from quantum complexity theory ruling out information theoretic secure, classical-client, blind quantum computation protocols.⁴ To achieve a fully classical client, one should weaken some of the assumptions: either allow some (well defined) leakage of information or aim for post-quantum computationally secure protocols.^g

A protocol was developed in Ma-hadev³¹ where a fully classical client delegates a (generic) quantum computation, without leaking information on the input and output. This protocol was post-quantum computationally secure. The key element in the construction was a mechanism to use a classical ciphertext to apply a (generic) quantum gate conditional on the corresponding plaintext, without ever decrypting and without leaking any information.

A second approach is to construct a mechanism that mimics a quantum channel by having a classical client interact with a quantum server,¹⁵ again with the consequence that the resulting protocol is post-quantum computational secure. Depending on the specifics of the simulated quantum channel, this functionality could enable classical clients to use all the protocols given in this section.

An important consequence is that classical clients could use *verifiable* blind quantum computation protocols. Here the clients can test the correctness of the delegated blind quantum computation, a feature crucial for commercial use of the quantum cloud. Finally, providing means that a classical agent can confirm the validity of a generic quantum computation

^g Classical client protocols with multiple non-communicating quantum servers have been proposed,^{23,34} based on quantum non-locality. However, the noncommunication of the quantum servers cannot be ensured indefinitely and the privacy gets compromised when those servers, eventually, communicate.

is a topic of great importance, both theoretically and practically, in its own right irrespective of whether it is used in a cryptographic setting or not.²² Another method to achieve verification of quantum computation in the post-quantum computational security setting, which does not necessarily rely in hiding the computation, was proposed in Mahadev.³²

In Cojocaru et al.,¹⁵ a quantum channel is replaced by a functionality of delegated pseudo-secret random qubit generator. Communicating random (secret) qubits is the only quantum communication required in many protocols (for example, Broadbent et al.¹² and Fitzsimons et al.²⁰). The key idea to achieve this functionality is to instruct the server to generate a state where some qubits are entangled, while some are unentangled. This is done in such a way that the connectivity is known to the client (that has access to trapdoor information) but is unknown to the server (that does not have access). The client uses this advantage and instructs the server to prepare an output qubit in a random state, of which the client knows its classical description while the server is totally ignorant. This exactly mimics a random single-qubit quantum channel.

The Future

The ability to communicate securely and compute efficiently is more important than ever to society. The Internet and increasingly the Internet of Things, has had a revolutionary impact on our world. Over the next 5–10 years, we will see a flux of new possibilities, as quantum technologies become part of this mainstream computing and communicating landscape. Future networks will certainly consist of both classical and quantum devices and links, some of which are expected to be dishonest, with functionalities of various sophistication, ranging from simple routers to servers executing universal quantum algorithms (see Figure 3). The realization of such a complex network of classical and quantum communication must rely on a solid novel foundation that, nevertheless, is able to foresee and handle the intricacies of real-life implementations and novel applications.

While post-quantum security paves the way for our classical Internet to remain safe in that era, quantum enhanced security aims to benefit actively from the development of quantum Internet in order to achieve unparalleled performances that are provably impossible using classical communication. Meanwhile quantum cloud services with various capabilities are becoming available. Quantum enabled security provides the platform that will ensure potential users that this new, unprecedented computational power in the quantum cloud, comes with the appropriate standards of accuracy, reliability and privacy. □

References

2013. Stored Encrypted Emails; <https://www.technewsworld.com/story/79117.html>.
2017. Quantum Internet Alliance; <http://quantum-internet.team/>.
2018. Google aims for quantum supremacy; <https://physicsworld.com/a/google-aims-for-quantum-supremacy/>.
4. Aaronson, S., Cojocaru, A., Gheorghiu, A. and Kashefi, E. On the implausibility of classical client blind quantum computing, 2017; arXiv:1704.08482.
5. Belovs, A. et al. Provably secure key establishment against quantum adversaries. In *Proceedings of the 12th Conf. Theory of Quantum Computation, Communication and Cryptography*. M.M. Wilde, ed. Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, 2018, 3:1–3:17.
6. Bennett, C.H. and Brassard, G. Quantum cryptography: Public key distribution and coin tossing. *Theor. Comput. Sci.* 560, P1 (2014), 7–11.
7. Bernstein, D.J. Introduction to post-quantum cryptography. *Postquantum Cryptography*. Springer, 2009, 1–14.
8. Boneh, D. et al. Random oracles in a quantum world. *Advances in Cryptology—ASIACRYPT 2011*. D.H Lee and X. Wang, eds. Springer.
9. Boneh, D. and Zhandry, M. Secure signatures and chosen ciphertext security in a quantum computing world. *Advances in Cryptology—CRYPTO 2013*. Springer, 361–379.
10. Brassard, G., Lütkenhaus, N., Mor, T. and Sanders, B.C. Limitations on practical quantum cryptography. *Physical Review Letters* 85, 6 (2000).
11. Braunstein, S.L. and Pirandola, S. Side-channel-free quantum key distribution. *Physical Review Letters* 108, 13 (2012), 130502.
12. Broadbent, A., Fitzsimons, J. and Kashefi, E. Universal blind quantum computation. In *Proceedings of the 50th Annual Symp. Foundations of Computer Science*. IEEE CS, 2009, 517–526.
13. Broadbent, A. and Schaffner, C. Quantum cryptography beyond quantum key distribution. *Designs, Codes and Cryptography* 78, 1 (2016), 351–382.
14. Buhrman, H. et al. Position-based quantum cryptography: Impossibility and constructions. *SIAM J. Comput.* 43, 1 (2014), 150–178.
15. Cojocaru, A., Colisson, L., Kashefi, E. and Wallden, P. On the possibility of classical client blind quantum computing, 2018; arXiv:1802.08759.
16. Dowling, J.P. and Milburn, G.J. Quantum technology: The second quantum revolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 361, 1809 (2003), 1655–1674.
17. Ekert, A. and Renner, R. The ultimate physical limits of privacy. *Nature* 507, 7493 (2014), 443.
18. Faugere, J.C. et al. Fast Quantum Algorithm for Solving Multivariate Quadratic Equations, 2017; arXiv:1712.07211.
19. Fitzsimons, J.F. Private quantum computation: An introduction to blind quantum computing and related protocols. *NJP Quantum Information* 3, 1 (2017), 23.
20. Fitzsimons, J.F. and Kashefi, E. Unconditionally verifiable blind quantum computation. *Physical Review A* 96 (2017), 012303.
21. Gagliardoni, T., Hüsing, A. and Schaffner, C. Semantic security and indistinguishability in the quantum world. *Advances in Cryptology—CRYPTO 2016*. M. Robshaw and J. Katz, eds. Springer, 60–89.
22. Gheorghiu, A., Kapourniotis, T. and Kashefi, E. Verification of quantum computation: An overview of existing approaches. *Theory of Computing Systems* (Jul 6, 2018).
23. Gheorghiu, A., Kashefi, E. and Wallden, P. Robustness and device independence of verifiable blind quantum computing. *New J. Physics* 17, 8 (2015), 083040.
24. Hanneke, D., Fogwell, S. and Gabrielse, G. New measurement of the electron magnetic moment and the fine structure constant. *Physical Review Letters* 100, 12 (2008), 120801.
25. Kaplan, M., Leurent, G., Leverrier, A. and Naya-Plasencia, M. Breaking symmetric cryptosystems using quantum period finding. *Advances in Cryptology—CRYPTO 2016*. M. Robshaw and J. Katz, eds. Springer, 207–237.
26. Kashefi, E., Music, L. and Wallden, P. The Quantum Cut-and-Choose Technique and Quantum Two-Party Computation, 2017; arXiv:1703.03754 (2017).
27. Liao, S.K. et al. Satellite-relayed intercontinental quantum network. *Physical Review Letters* 120, 3 (2018), 030501.
28. Lindell, Y. How to simulate it—A tutorial on the simulation proof technique. *Tutorials on the Foundations of Cryptography*. Springer, 2017, 277–346.
29. Lo, H.K., Curty, M. and Qi, B. Measurement-device-independent quantum key distribution. *Physical Review Letters* 108, 13 (2012), 130503.
30. Lydersen, L. et al. Hacking commercial quantum cryptography systems by tailored bright illumination. *Nature Photonics* 4, 10 (2010), 686.
31. Mahadev, U. Classical homomorphic encryption for quantum circuits. In *Proceedings of the IEEE 59th Annual Symposium on Foundations of Computer Science* (Paris, France, 2018), 332–338.
32. Mahadev, U. Classical verification of quantum computations. In *Proceedings of the IEEE 59th Annual Symposium on Foundations of Computer Science* (Paris, France, 2018), 259–267.
33. Bohr, N. On the constitution of atoms and molecules. *The London, Edinburgh, and Dublin Philosophical Magazine and J. Science* 26, 151 (1913), 1–25.
34. Reichardt, B.W., Unger, F. and Vazirani, U. Classical command of quantum systems. *Nature* 496, 7446 (2013), 456.
35. Shor, P.W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Review* 41, 2 (1999), 303–332.
36. Unruh, D. Quantum proofs of knowledge. *Advances in Cryptology—EUROCRYPT 2012*. D. Pointcheval and T. Johansson, eds. Springer, 135–152.
37. van de Graaf, J. Towards a Formal Definition of Security for Quantum Protocols. Ph.D. Dissertation, 1998. Montreal, Canada.
38. Von Neumann, J. *Mathematical Foundations of Quantum Mechanics*. Number 2. Princeton University Press, 1955.
39. Watrous, J. Zero-knowledge against quantum attacks. *SIAM J. Comput.* 39, 1 (2009), 25–58.
40. Xu, F. et al. Experimental quantum fingerprinting with weak coherent pulses. *Nature Communications*, (2015), 8735.

Petros Walden (petros.walden@ed.ac.uk) is Lecturer at the University of Edinburgh, Scotland, U.K.

Elham Kashefi (ekashefi@staffmail.ed.ac.uk) is a professor at the University of Edinburgh, Scotland, U.K. and Sorbonne Université, CNRS, Laboratoire d’Informatique de Paris, France.

Copyright held by authors/owners.



Watch the author discuss this work in the exclusive *Communications* video. <https://cacm.acm.org/videos/cyber-security-in-the-quantum-era>

ACM Welcomes the Colleges and Universities Participating in ACM's Academic Department Membership Program

ACM now offers an Academic Department Membership option, which allows universities and colleges to provide ACM Professional Membership to their faculty at a greatly reduced collective cost.

The following institutions currently participate in ACM's Academic Department Membership program:

- AAU Klagenfurt
- Abilene Christian University
- Amherst College
- Appalachian State University
- Armstrong State University
- Ball State University
- Bellevue College
- Berea College
- Binghamton University
- Boise State University
- Bryant University
- Calvin College
- Colgate University
- Colorado School of Mines
- Cornell University
- Creighton University
- Cuyahoga Community College
- Edgewood College
- Franklin University
- Gallaudet University
- Georgia Institute of Technology
- Governors State University
- Harding University
- Harvard University
- Hofstra University
- Hope College
- Howard Payne University
- Indiana University - Bloomington
- Indiana University Bloomington, Information & Library Science
- Kent State University
- La Sierra University
- Messiah College
- Metropolitan State University
- Missouri State University
- Montclair University
- Mount Holyoke College
- New Jersey Institute of Technology
- Northeastern University
- Old Dominion University
- Pacific Lutheran University
- Potomac State College of West Virginia
- Regis University
- Roosevelt University
- Rutgers University
- SUNY Oswego
- Saint Louis University
- San Jose State University | Davidson College of Engineering
- Shippensburg University
- Simmons College
- St. John's University
- Stanford University
- Stetson University
- The Ohio State University
- The Pennsylvania State University
- The State University of New York at Fredonia
- The University of Alabama
- The University of Memphis
- Trine University
- Trinity University
- UC San Diego
- UNC Charlotte
- USC, University of Southern California
- Union College
- Union University
- University of California, Riverside
- University of California, Santa Cruz
- University of Colorado Boulder
- University of Colorado, Denver
- University of Houston
- University of Illinois at Chicago
- University of Jamestown
- University of Liechtenstein
- University of Maryland, Baltimore County
- University of Nebraska at Kearney
- University of Nebraska at Omaha
- University of New Mexico
- University of North Dakota
- University of Pittsburgh
- University of Porto, Faculty of Engineering
- University of Puget Sound
- University of Victoria
- University of Wisconsin - Parkside
- University of Wyoming
- University of the Fraser Valley
- Virginia Commonwealth University
- Wake Forest University
- Wayne State University
- Wellesley College
- Western New England University
- William Jessup University - Rocklin Campus
- Worcester State University

Through this program, each faculty member receives all the benefits of individual professional membership, including *Communications of the ACM*, member rates to attend ACM Special Interest Group conferences, member subscription rates to ACM journals, and much more.

For more information: www.acm.org/academic-dept-membership



Association for
Computing Machinery

research highlights

P. 132

**Technical
Perspective**

**Was Edgar Allan Poe
Wrong After All?**

By Gilles Brassard

P. 133

**Fully Device Independent
Quantum Key Distribution**

By Umesh Vazirani and Thomas Vidick



Technical Perspective

Was Edgar Allan Poe Wrong After All?

By Gilles Brassard

FOR THOUSANDS OF years, the battle has been raging between codemakers and codebreakers. There have been dramatic reversals of fortune throughout history, sometimes with spectacular consequences that have changed the course of civilization. For instance, the world today would be a very different place had the Allies not cracked the German's Enigma cipher during the Second World War. Numerous popular books have been written about the war of codes, such as Simon Singh's bestseller, *The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography*. Today, our entire economy, and the very existence of cyberspace, depends crucially on our ability to communicate confidentially. But can we really?

The billion-dollar question is: Will it be codemakers or codebreakers that emerge in the end as undisputed victors? Or perhaps the game of cat-and-mouse will go on forever, with codemakers inventing evermore clever encryption schemes, only to be defeated by even cleverer codebreakers. In 1841, poet and novelist Edgar Allan Poe famously wrote in the improbably named *Graham's Lady's and Gentleman's Magazine* that "It may be roundly asserted that human ingenuity cannot concoct a cipher which human ingenuity cannot resolve." This was quite a bold statement because Blaise de Vigenère had published in 1585 a scheme then known as *le chiffre indéchiffrable*, which nobody had yet been able to break. How could Poe have foreseen that it would be broken by Charles Babbage just a few years later? So, perhaps he was right after all.

Then computers came into the game and considerably more complex ciphers were designed. Somehow related to the $P \neq NP$ conjecture, it appeared the emergence of increasingly powerful computers was to the advantage of codemakers. The in-

vention of public-key cryptography in the 1970s seemed to turn the tide resolutely in favor of codemakers. In particular, the publication of the RSA cryptosystem in these pages (CACM Feb. 1978) was a turning point. Poe, it would seem, was wrong.

That was until Peter Shor discovered in 1994 that a quantum computer could efficiently break not only RSA but also the earlier Diffie-Hellman key establishment scheme (even based on elliptic curves), essentially bringing to its knees the entire cryptographic infrastructure on which the presumed security of the Internet currently rests. Not to worry: quantum computers seemed to be a threat for the far future in 1994 ... but not anymore. Poe, it would seem, would have the last laugh.

But already in 1983, Charles Bennett and I had invented *quantum key distribution* (QKD), thus harnessing the power of the quantum for the benefit of codemakers one decade before Shor harnessed it to serve the opposite side. As Asher Peres once said, "The quantum taketh away and the quantum giveth back." Furthermore, the implementation of QKD seemed so much easier than the construction of a full-scale quantum computer necessary to run Shor's algorithm. In 1996, the unconditional security of QKD was proved by Dominic Mayers (even against quantum computers) and increasingly sophisticated prototypes were being built. Hence the issue was finally settled: Poe was wrong.

That was until *Quantum Hackers* emerged, many of whom under the leadership of Vadim Makarov. They realized the security of QKD was unconditional only if the apparatus could be built exactly as defined in the theoretical papers, a seemingly impossible task. In 2009, they demonstrated a full-field implementation of a complete attack on a running

QKD connection. The specific weakness thus uncovered by Makarov was quickly patched, but another emerged just as quickly. Poe's statement could be paraphrased as "It may be roundly asserted that engineers ingenuity cannot implement a QKD apparatus without leaving a loophole which engineers ingenuity can exploit." Thus, the game of cat-and-mouse would continue forever, albeit lifted from the realm of mathematics to that of engineering.

This is where the following paper by Umesh Vazirani and Thomas Vidick enters the stage. Artur Ekert realized as early as 1991 that a different kind of quantum cryptography was possible by harnessing *entanglement*, which is arguably the most nonclassical manifestation of quantum theory. Even though Ekert's original protocol did not offer any security above and beyond my earlier invention with Bennett, he had planted the seed for a revolution. It was realized by several researchers in the mid-2000s that entanglement-based protocols could lead to unconditional security even if they are imperfectly implemented—even if the QKD apparatus is built by the eavesdropper, some argued. For a decade, these purely theoretical ideas remained elusive and seemed to require unreasonable hardware, such as an apparatus the size of the galaxy! Vazirani and Vidick's paper provides an unexpectedly simple and elegant solution, indeed one that is almost within reach of current technology. Once it becomes reality, codemakers will have won the definitive battle, Poe's prophecy notwithstanding. □

Gilles Brassard is a professor at Université de Montréal, Quebec, Canada, and Canada Research Chair in Quantum Information Science.

Copyright held by author.

Fully Device Independent Quantum Key Distribution

By Umesh Vazirani and Thomas Vidick

Abstract

Quantum cryptography promises levels of security that are impossible to attain in a classical world. Can this security be guaranteed to classical users of a quantum protocol, who may not even trust the quantum devices used to implement the protocol?

This central question dates back to the early 1990s when the challenge of achieving Device-Independent Quantum Key Distribution (DIQKD) was first formulated. We answer the challenge by rigorously proving the device-independent security of an entanglement-based protocol building on Ekert's original proposal for quantum key distribution. The proof of security builds on techniques from the classical theory of pseudo-randomness to achieve a new quantitative understanding of the non-local nature of quantum correlations.

1. INTRODUCTION

The gold standard in classical cryptography is semantic security¹⁶—given an encoded message (ciphertext), any polynomial time adversary should have a negligible chance of obtaining any information whatsoever about the plaintext (the message that was encoded). Classical cryptographers have developed encryption schemes that are able to provide this guarantee by relying on un-proven assumptions about the computational difficulty of number-theoretic problems, such as factoring large numbers,²⁸ or other more general complexity-theoretic assumptions. Such cryptographic assumptions are stronger than $P \neq NP$, and considered unlikely to be proven in the foreseeable future, so one may ask whether any cryptosystems can have security guarantees that do not rely on unproven computational assumptions.

Bennett and Brassard's seminal discovery⁷ of a Quantum Protocol for Key Distribution (QKD) in 1984 opened up the possibility for a different kind of security—"unconditional security," or security based solely on the validity of universally accepted physical laws. QKD is a protocol that allows two distant parties to collaboratively and privately generate a perfectly random string, called the users' key. Once the users have generated a private shared key they can use it to communicate with perfect security as follows. To send a message securely, the sender uses the shared key as one time pad: the encoded message is simply the bit-wise XOR (parity) of the key with the message. The semantic security of this scheme is easy to verify, provided the one-time pad (the key) is completely random and unpredictable to any adversary.

The first rigorous proofs of security of QKD,^{21,31} established over two decades ago, appeared to guarantee this ultimate notion of security could be achieved. However, the promise was short-lived. QKD researchers soon realized that practical implementations fell prey to types of attacks that were

not accounted for by the security proofs, including side-channel attacks⁸ and photon receptor blinding attacks.²⁰ More generally, these attacks pointed to a fundamental issue: given the remarkable features of quantum mechanics, such as superpositions, entanglement, and the uncertainty principle—some of which made QKD possible in the first place—how can one trust that there are not novel ways of attacking the quantum devices implementing a QKD protocol, unbeknownst to the honest, but classical, users of the protocol? The basic assumption that the user has perfect control over and trust in her quantum devices in a cryptographic protocol, on which all security proofs crucially relied, now appeared wholly unrealistic.

In a paper which appeared in the proceedings of FOCS'98, Mayers and Yao²² proposed a first principles approach to this question in the form of a new security paradigm called *device independence*.^a In this formulation, all quantum devices used in a protocol are modeled as black boxes with classical inputs and outputs. The user's confidence in a quantum device should be based only on the observed input-output behavior of the device. Concretely, in the context of QKD, a secure protocol should include a practical test that guarantees that the users' quantum devices behave according to specification, even in the scenario where the devices may have been manufactured by an adversarial party. More precisely, any execution of the protocol should either abort, or contain a proof, based only on the classical statistics observed during the execution, that no eavesdropper (a term we will use synonymously with "adversary") may have obtained any partial information about the final key (except with exponentially small probability).

Device independence may appear to set an impossibly high standard. Yet already a few years before the notion was introduced a different protocol for QKD had been proposed by Ekert,¹⁴ together with intuition suggesting that it should have strong security guarantees reminiscent of device independence. The starting point for Ekert's protocol was provided by the properties of a remarkable quantum phenomenon, entanglement—the very phenomenon which so famously puzzled the authors of the "EPR paradox".¹³ A canonical example of entanglement is the Bell state on two quantum bits, or *qubits*: $1/\sqrt{2}|00\rangle + 1/\sqrt{2}|11\rangle$. Physically, such a maximally entangled pair of qubits can be realized in the polarization of a pair of photons, or by any pair of spin- $\frac{1}{2}$ particles, such as electrons.

^a The term "device independence" was only introduced much later, in 2007.²⁴

A technical version of this paper was published in *Physical Review Letters*, Sept. 29, 2014.

In pioneering work in the 1960s (see Bell⁶) the physicist John Bell showed that suitable measurements on the two qubits in a Bell state generate correlations between their outputs that cannot be reproduced by non-communicating classical devices. Such non-local correlations provide a “test of quantumness,” one that distinguishes quantum mechanics from the kinds of hidden variable theories that Einstein championed. Non-local correlations of this form have been experimentally verified to exquisite precision in the form of Bell inequality violations.^{15, 18, 30}

The simplest example of a Bell experiment was discovered by Clauser et al.⁹ The experiment can be formulated as a game, the CHSH game, named after its inventors Clauser, Horne, Shimony, and Holt that plays an essential role in our protocol. In this game, devices D_A and D_B are given as input uniformly random bits, x and y respectively. Their goal is to return bits, a and b respectively, such that $xy = a \oplus b$. An easy argument shows that if D_A and D_B are classical, that is, each device’s output is computed as a function of the device’s input only, they cannot succeed with probability greater than 3/4, even if the devices are allowed to access a common shared random string. By contrast, if D_A and D_B are quantum, each containing one qubit of a Bell state, then by performing suitable quantum measurements on their qubits they can achieve a success probability of $\cos^2 \pi/8 \approx 0.85 > 3/4$.

The importance of non-local correlations for device independence lies in the intuition that under certain conditions these correlations, which can be tested classically via a Bell experiment, can pin down the entangled state shared by the two devices—a phenomenon known as rigidity—as well as preclude entanglement with a third party such as the Eavesdropper—a phenomenon known as monogamy. Unfortunately, beyond its intuitive formulation monogamy is notoriously difficult to quantify, and this has been one of the main obstacles to a formal proof of security in the device-independent framework.

A significant step was taken by Barrett et al.,⁵ who were able to show security of a protocol generating a single bit of key based only on the no-signaling assumption that no information is exchanged between the users’ devices, or between the devices and the eavesdropper. This is a weaker assumption than the quantum formalism, which implies non-signaling but is generally more restrictive. Their work inspired a large number of follow-up papers^{1, 2, 29} which sought security for protocols generating multiple bits, and attempted to leverage the quantum formalism to obtain a more fine-grained security analysis, leading to potential improvements in key rates. Unfortunately all these attempts ran into a fundamental obstacle, which is that one cannot use Bayes’ rule for conditional probability in a quantum setting: the Eavesdropper’s maximum probability of guessing two bits of the key $k_1 k_2$ cannot be expressed as the product of her maximum probability of guessing k_1 times her maximum probability of guessing k_2 given that she guessed k_1 . This has to do with the quantum nature of the Eavesdropper, and the fact that different measurements on the same quantum state are not always compatible (they may not commute).

A complete proof of security of a variant of Ekert’s protocol was first given in previous work of the authors.³⁵ The goal of this paper is to highlight two key components of this security proof. The first is the use of deep properties of

randomness extractors (extremal objects from the computational complexity-based theory of pseudorandomness) and their quantum-proof extensions to bypass obstacles, such as the absence of a suitable Bayes’ rule more specifically, for analyzing quantum protocols generating multiple bits. Building upon previous work on certifiable quantum random number generation,³⁴ the proof of security uses an extraction-based proof technique called the “quantum reconstruction paradigm” to perform a reduction from global attacks of the Eavesdropper to attacks on a single round of the protocol. This component of the proof should be of broad interest to complexity theorists interested in randomness extractors and multi-player games, and the exposition tries to make these aspects accessible.

The second component of the proof rules out single-round attacks by the Eavesdropper. This involves arguing that any such attack would imply an impossibly good strategy for three players involved in a simple non-local “guessing game,” that would contradict the non-signaling principle. The proof takes the form of a reduction between guessing games, and may be particularly relevant to cryptographers in view of the recent interest in the properties of no-signaling distributions in the context of delegated computation.¹⁹

Our security proof operates within a considerable generalization of the framework of Bell experiments that encompasses the study of correlations between the outputs of three quantum devices (the third device being used to model an Eavesdropper to the protocol) involved in a complex interaction which includes some limited communication between the devices. This generalization captures new physical phenomena, such as the monogamy of entanglement mentioned earlier. In our exposition we aim to abstract as much of the quantum formalism as possible by taking the point of view of classical users in the protocol, thereby also highlighting the classically understandable elements of the proof of security while postponing as far as possible any knowledge about the quantum formalism required to provide a low-level modeling of the quantum devices used.

To achieve device independent security in an experiment, it suffices to implement devices that can carry out the honest party’s protocol with sufficient fidelity that their input-output behavior passes the statistical test specified in the protocol. This is in contrast to traditional security proofs that require e.g. a device that is guaranteed to emit a single qubit at a time, a criterion that cannot be verified based on statistical data. The recent experimental demonstrations of Bell tests reproducing the desired correlations even under stringent adversarial assumptions (so-called “loophole-free”)^{5, 18, 30} make it a near-certainty that device-independent protocols for QKD will soon be realized. An even bigger canvas for such experiments is provided by the recent demonstration of the feasibility of distributing entangled qubits from satellite to distant ground stations.³⁶ These are exciting times for quantum cryptography!

2. QUANTUM STATES, ENTANGLEMENT AND THE CHSH GAME

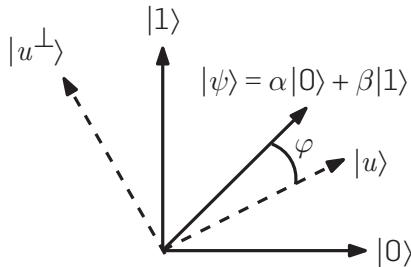
The qubit is the fundamental unit of quantum information. For our purposes it suffices to consider the following simplified definition: a qubit is a particle that can be in a real

superposition of two states, labeled $|0\rangle$ and $|1\rangle$. The superposition is represented by a unit vector in the two dimensional real plane, and may be specified by a parameter θ as: $(\cos \theta \sin \theta)^T$, or in ket notation as $|\psi_\theta\rangle = \cos \theta |0\rangle + \sin \theta |1\rangle$. This means that a unit vector pointing in the x direction indicates the state $|0\rangle$, and a unit vector pointing in the y direction represents the state $|1\rangle$ ^b.

An observation, or measurement, performed on a qubit is specified by a choice of basis. The distinguished basis formed by the x and y -axes is called the standard basis. The Born rule for the outcomes of quantum measurements specifies that a measurement of the state $|\psi_\theta\rangle$ in the standard basis yields the outcome 0 with probability $\cos^2 \theta$, and 1 with probability $\sin^2 \theta$. Once the measurement has been performed, the qubit collapses to the post-measurement state consistent with the outcome obtained, $|\psi_0\rangle = |0\rangle$ or $|\psi_{\pi/2}\rangle = |1\rangle$ respectively. More generally, for any angle φ the qubit can also be measured in the basis obtained by rotating the standard basis by an angle of φ . Such a measurement yields the outcome 0 with probability $\cos^2(\theta - \varphi)$, and 1 with probability $\sin^2(\theta - \varphi)$ (See Figure 1).^c The post-measurement states are $|\psi_\varphi\rangle$ or $|\psi_{\pi/2+\varphi}\rangle$ respectively.

Now we turn to a discussion of entanglement. The most fundamental entangled state is the Bell state, which is a maximally entangled state of two qubits. Using the ket notation the Bell state is written $|\phi^+\rangle = \frac{1}{\sqrt{2}}|0\rangle|0\rangle + \frac{1}{\sqrt{2}}|1\rangle|1\rangle$. To understand the remarkable properties of this state, let us start by describing the result of measuring both qubits in the same basis ($|\psi_\varphi\rangle$, $|\psi_{\pi/2+\varphi}\rangle$): first note that due to symmetry, the outcome of the measurement on each qubit is uniformly random; that is, 0 or 1, with probability 1/2 each. But the symmetry of the Bell state goes deeper. Note that the Bell state is rotationally invariant and can be written as $|\phi^+\rangle = \frac{1}{\sqrt{2}} |\psi_\varphi\rangle |\psi_\varphi\rangle + \frac{1}{\sqrt{2}} |\psi_{\varphi+\pi/2}\rangle |\psi_{\varphi+\pi/2}\rangle$ (this can be easily verified by direct calculation). This means that the outcome of a measurement on the two qubits is that both outcomes are 0, or both outcomes are 1, each with probability 1/2. Moreover the post-measurement state is $|\psi_\varphi\rangle|\psi_\varphi\rangle$, or $|\psi_{\pi/2+\varphi}\rangle|\psi_{\pi/2+\varphi}\rangle$ respectively. This conclusion

Figure 1. Measuring ψ in the standard basis yields the outcome 0 with probability $|\alpha|^2$. Measuring it in the basis (u, u^\perp) yields the outcome u with probability $\cos^2 \varphi$.



^b An example of a no-signaling correlation that is neither local nor quantum is the family of distributions $p(a, b|x, y) = 1/2$ if and only if $a \oplus b = xy$, for $a, b, x, y \in \{0, 1\}$. This family gives a success probability 1 in the CHSH game with probability 1.

^c Note here we use the notation “0” to designate the outcome, even though the associated basis element is the vector $|\psi_\varphi\rangle$, not $|0\rangle$. This is a standard convention; labels for outcomes are arbitrary and have no physical meaning.

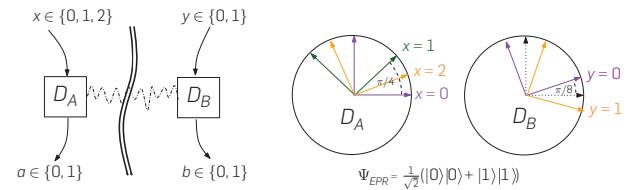
holds regardless of the spatial separation between the two qubits, a conclusion that appears very counterintuitive at first glance: how does the second qubit “know” the basis in which the first qubit has been measured, and the outcome obtained?

Such natural puzzlement, however, is an artifact of our presentation (the same given by Einstein, Podolsky and Rosen!). Indeed, a simple classical scenario recreates the same phenomenon. Consider two coins with the same random orientation, so that the normal to their surface makes the same (random) angle θ in the $x - y$ plane. The two coins are then spatially separated while leaving their orientation unchanged and measured. By measurement, we mean pick an angle φ , look at the coin along the angle φ (looking from infinity, towards the coin), and report whether one sees a heads or tails (0 or 1). Clearly the outcome is uniformly random since the orientation θ was chosen uniformly at random. And if we observe (measure) both coins from the same angle φ , the outcomes will be identical.

The fundamental distinction between the quantum and classical settings manifests itself when the two qubits are measured in different bases. If we measure the first qubit in the standard basis and the second in a φ -rotated basis, then as previously each outcome (0 or 1) is individually random, but the measurement rule indicates that the two outcomes will match with probability $\cos^2 \varphi$. This differs from the statistics obtained in the classical scenario described above, in which the chance of obtaining matching outcomes is $|1 - \varphi/\pi|$ for $\varphi \in [0, \pi]$. The difference between these two numbers, φ^2 vs. φ/π for small φ , is the foundation for Bell’s test of quantumness.

We can now see how this difference plays out in a particularly elegant way in the case of the CHSH game. The quantum strategy starts with Alice and Bob sharing a Bell state. Each of them performs a measurement of their qubit in a basis which depends on their input bit (x or y), and announces the result: Alice measures her qubit in a θ_A -rotated basis, with $\theta_A = \frac{\pi}{4}x$, and Bob measures his qubit in a θ_B -rotated basis, with $\theta_B = \frac{\pi}{8} - \frac{\pi}{4}y$ (Figure 2). In the three cases where $xy = 0$, the players measure in bases that are $\pi/8$ -rotated relative to each other, and therefore output the same bit (i.e. $a \oplus b = 0$) with probability $\cos^2 \pi/8$. In the case where $xy = 1$, they measure in bases that are $3\pi/8$ -rotated relative to each other, and therefore output different bits (i.e. $a \oplus b = 1$) with probability $\cos^2 \pi/8$ as well. Thus in each of the four cases they succeed with probability exactly $\cos^2 \pi/8$. In contrast, the classical strategy

Figure 2. Our proof of security relies on one of the most fundamental Bell inequalities, the CHSH inequality,⁹ which is pictured here as a small game. Honest devices measuring a Bell pair in the bases indicated on the right of the figure will produce outputs such that $\Pr(a \oplus b = xy) = \cos^2 \pi/8$ whenever $x, y \in \{0, 1\}$, and $a = b$ whenever $x = 2$ and $y = 0$.



based on the use of random coins described earlier, using the same angles as the quantum strategy (the angles should be doubled to account for the different way these angles are used in quantum or classical strategies), would produce valid outcomes with probability $1 - 1/4 = 3/4$ in each case.

3. EKERT'S PROTOCOL AND A PROVABLE VARIANT

We are ready to move on to the more complex task of key distribution. The goal of this task is for two trusted but distant users Alice and Bob to agree on a shared n -bit key $K \in \{0, 1\}^n$. A key distribution protocol must guarantee that if the protocol runs to completion then the users produce identical keys that are indistinguishable from a uniformly random string to any eavesdropper. The only resources available to the users are trusted local random number generators and a public, authenticated quantum communication channel.

3.1. Ekert's protocol

We first introduce a slight variant of Ekert's protocol, that carries the same intuition for security. The users, Alice and Bob, initially share a large number, N , of Bell states. These states could have been generated by Alice, with one qubit from each state transmitted to Bob. Since all communication channels are public, an eavesdropper may have interfered with the qubits. Taking an adversarial point of view, it is convenient to consider a symmetric formulation, where Alice and Bob have been provided with quantum devices, D_A and D_B , by some untrusted provider. The devices contain an arbitrary state, on which they perform arbitrary measurements when operated by the user.

In the protocol, the devices D_A and D_B are used N times in sequence. At each use, Alice's device D_A can take one of three possible inputs $x_i \in \{0, 1, 2\}$, and Bob's device D_B has two inputs $y_i \in \{0, 1\}$. These inputs are chosen uniformly at random by the users in each round. The ideal prescription for the devices is such that on two of Alice's choices of input (and both Bob's inputs) the device is supposed to perform measurements that follow the optimal quantum strategy for the CHSH game (Figure 2). This means that with probability 2/3, Alice and Bob use their devices to play the CHSH game, in which case they check the required correlations satisfy the CHSH game condition, with sufficiently high frequency on average over the number of such game rounds. The third input for Alice's device is meant to indicate a measurement that is identical to Bob's device's measurement on input 0. This means that with probability $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$ Alice and Bob choose this matching pair of inputs, in which case they expect to obtain identical outcomes from which the key can be generated.

Ekert's intuitive argument for security of this protocol relied on a few observations. The first is that the eavesdropper cannot obtain any meaningful information while the qubits are in transit from the source, because there is no information encoded in the Bell state: as formulated in Ref. Ekert,¹⁴ the information only "comes into being" when the qubits are measured by Alice and Bob. Therefore the only reasonable attack for the Eavesdropper at this stage is to keep a quantum state that may be partially or fully entangled with the qubits that are present in Alice and Bob's devices. A second observation is

that the CHSH game test can be used to build confidence that the qubits are indeed in a Bell state. Intuitively, the Bell state is a pure state, which by definition cannot be entangled with any qubits held by the eavesdropper. Therefore, even after the users have interacted with their devices there should be no additional information that the eavesdropper may gain by performing measurements on her system. Ekert did not provide a proof of security for his QKD protocol. A proof in the weaker model for security, in which the users' quantum devices are trusted, was first achieved by reduction to the BB'84 protocol.³¹ The proof bypassed Ekert's original intuition, instead building on the nascent theory of quantum error-correction and its application to entanglement distillation, and it crucially relied on the assumption that the quantum devices are fully trusted.

3.2. Protocol E1

We will show security for a variant of the protocol described in the previous section, described in Figure 3. As previously, the users, Alice and Bob, are provided with arbitrary quantum devices D_A , which takes as input a trit $x_i \in \{0, 1, 2\}$ and D_B , which takes as input a bit $y_i \in \{0, 1\}$. The devices are used sequentially in N rounds, each time returning an output bit, $a_i, b_i \in \{0, 1\}$ respectively. The ideal specification for the devices is the same as in the previous protocol.

We divide the protocol into three phases. The first phase is the *generation phase*, in which the users execute their respective device N times in sequence by making uniformly random choices of inputs and collecting the N output bits. Then begins the *testing phase*: the users publicly select a constant fraction γ of the rounds uniformly at random and evaluate the devices' success probability in a variant of the CHSH game on average over the input-output pairs generated in those rounds. If they observe that the success probability deviates from the optimum by more than the allowed tolerance η , the users abort

Figure 3. Our DIQKD protocol, protocol E1.

1. Inputs: N = number of rounds, γ = testing fraction, η = noise tolerance.
2. *Generation:* For rounds $i = 1, \dots, N$:
Alice selects $x_i \in \{0, 1, 2\}$, and Bob selects $y_i \in \{0, 1\}$, uniformly at random. They input x_i, y_i into their device, obtaining outputs $a_i, b_i \in \{0, 1\}$ respectively.
3. *Testing:* Alice chooses a random subset $\mathbf{B} \subseteq \{1, \dots, N\}$ of test rounds of size $\lceil \gamma N \rceil$ and shares it publicly with Bob. Alice and Bob announce their input/output pairs in all rounds in \mathbf{B} . They compute the fraction of pairs in \mathbf{B} that satisfy the CHSH condition $a_i \oplus b_i = x_i y_i$. If this fraction is smaller than $\cos^2 \frac{\pi}{8} - \eta$ they abort the protocol. Similarly, they compute the fraction of pairs such that $a_i = b_i$, conditioned on $(x_i, y_i) = (2, 0)$, and abort if the fraction is smaller than $1 - \eta$.
4. *Extraction:* Alice and Bob publicly reveal their N choices of inputs. Let \mathbf{C} be the set of check rounds: rounds i for which $(x_i, y_i) = (2, 0)$. They perform information reconciliation on the rounds in \mathbf{C} , followed by privacy amplification, to obtain their respective key.

the protocol. Otherwise they proceed to the *extraction phase*. For this they exchange their inputs in all rounds and discard rounds in which the inputs were not $(x_i, y_i) = (2, 0)$. Outputs from the remaining rounds, called the check rounds C, are kept private by each user and designated as the *raw key*, K_A and K_B respectively. Based on these strings the users perform post-processing steps of error reconciliation and privacy amplification to obtain the final shared key K . Information reconciliation is a procedure, based on error-correcting codes, which aims to ensure that $K_A = K_B$ after a small correction performed publicly. We will not discuss this standard procedure here. The goal of privacy amplification is to amplify the secrecy present in the raw key, from partially unknown to the Eavesdropper, to indistinguishable from uniform from the point of view of the Eavesdropper. We will describe this procedure in detail later.

The most important point of departure from Ekert's protocol lies in the small fraction γ of rounds in which the CHSH game correlations are tested. Skimping on testing may appear to weaken the security of the protocol: why not use all possible correlations for the test? On the other hand, testing involves publicly revealing the input/output pairs for those rounds, and this runs the risk of dishonest devices leaking information to the eavesdropper. Indeed, as we will see the proof of security of the protocol will require us to set a small, constant upper bound on the fraction γ of rounds selected for testing.

We formalize the intuition underlying protocol E1 in terms of two concepts, neither of which had been clearly formulated at the time of Ekert's paper. The first is known as rigidity. This is the observation that there is an essentially unique way to achieve the optimal success probability of $\cos^2 \pi/8$ in the CHSH game: any optimal strategy for the devices D_A and D_B in the game is locally equivalent to the specific strategy based on the Bell state described earlier. The second concept is the monogamy of correlations. This idea builds on the first by stating that maximally entangled qubits are necessarily pure, thus cannot share their "maximally entangled degrees of freedom" with any other qubits. It is a formalization of Ekert's intuition that, by witnessing correlations that certify that their devices share a Bell state, the users effectively preclude the possibility of entanglement between the devices and any malicious eavesdropper.

One possible approach to a formal proof of security of the protocol would be to give a quantitative formulation of both concepts. While this can be done within the mathematical formalism of quantum mechanics, there are strong impediments to the use of the resulting theory for proving security in any practically meaningful sense. First of all, achieving device independence prevents us from making any a priori assumption on the quantum systems used, such as a bound on the dimension of the underlying Hilbert space. Without such bound it is not a priori obvious how the distance to the optimal strategy scales as a function of the deviation from optimality of the statistics observed, a difficulty already faced by Mayers and Yao.²² More importantly, the interaction scenario which takes place in the QKD protocol is complex and involves multiple sources of information leaked to the eavesdropper. While a proof of security along these lines was given in Ref. Reichardt et al.,²⁷ the bounds obtained are too weak to tolerate any errors in the execution of the protocol (such as unavoidable photon losses, false detection events, etc.).

4. OVERVIEW OF PROOF OF SECURITY

The proof of security can be decomposed into two parts. The first analyzes a single randomly chosen round of the protocol. It shows that if the protocol succeeds, then with high probability the devices' output in the chosen round must be at least partially unknown to the Eavesdropper. To do so, the challenge faced by the Eavesdropper is formulated in terms of a two player guessing game, and the information of the Eavesdropper is bounded by means of a reduction to a trivial guessing game. The maximum success probability in the latter game is bounded by virtue of the no-signaling principle, that is, in the absence of communication between the two players, the output distribution of one player must be independent of the other player's input. This part of the proof does not resort to the quantum formalism at all (though doing so can be used to give stronger quantitative bounds).

The second part of the proof shows that over a large number N of rounds, the protocol yields $\Omega(N)$ bits of shared random key that are secure against the eavesdropper. Analyzing multiple rounds of the protocol runs into fundamental obstacles having to do with the quantum nature of the Eavesdropper and her attack. Ideally we would like to associate the extraction of fewer than $\Omega(N)$ bits of shared random key with the failure of the guessing game argument in some round, resulting in a contradiction. The issue is that the Eavesdropper's attack could be a global measurement on her part of the quantum state based on the classical information exchanged in all rounds of the protocol, leading to a loss of control over the sequential nature of the protocol. We will get around these obstacles by appealing to powerful results from the theory of randomness extractors to directly deal with correlations between the classical information generated by Alice, Bob and the Eavesdropper. Roughly, this paradigm allows us to directly conclude that if the key generated by the protocol is distinguishable from random by the eavesdropper, then there is an (efficient) classical procedure to reconstruct Alice's raw key with non-negligible probability. This allows us to use classical information-theoretic tools to perform a reduction to the guessing game.

5. GUESSING GAMES

At its heart the security guarantees provided by a single round of protocol E1 hinge on the following guessing game (Figure 4). We start with an augmented variant of the CHSH game, where in addition to the CHSH game inputs Alice can be provided

Figure 4. The guessing game. Any devices satisfying both the CHSH condition $a \oplus b = xy$ and the guessing condition $c = a$ with high enough probability must allow signaling between D_A and D_B .

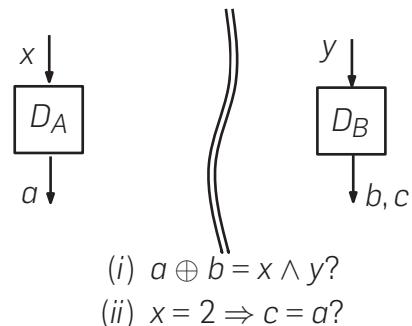
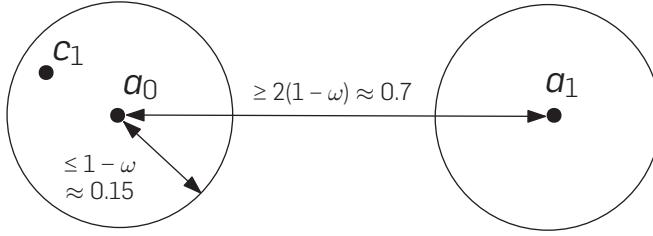


Figure 5. The proof of the guessing lemma. Using the winning condition for the CHSH game, wherever b_1 is the distance between a_0 and a_1 must be at least $2(1 - \omega)$. This implies the balls of radius $(1 - \omega)$ centered are a_0 and a_1 are disjoint.



with a third input, labeled 2 (each input is chosen with probability 1/3). If inputs to the players are $(x, y) = (2, 0)$ then their outputs should match. Assuming the no-signaling principle and the validity of quantum mechanics (specifically, that $\omega = \cos^2 \pi/8$ is the optimal bound in the CHSH game), it is easy to see that Alice and Bob can win the augmented CHSH game with probability $\frac{2}{3}\omega + \frac{1}{3}$, which is optimal.

Now we consider adding one more rule to the game, the “guessing rule”: Bob should always produce a second output, $c \in \{0, 1\}$. The guessing rule requires that c should match Alice’s output a whenever her input is $x = 2$ (if $x \in \{0, 1\}$ then there is no requirement on c). We will show that Alice and Bob cannot achieve close to the optimal probability of $\frac{2}{3}\omega + \frac{1}{3}$ in the extended CHSH game, and simultaneously satisfy the guessing rule on Bob’s output c with probability close to 1: whatever Bob does, he cannot collaborate with Alice to succeed in the CHSH game, while simultaneously producing a valid guess for her output. This is a manifestation of the phenomenon of monogamy of quantum correlations alluded to earlier. To understand the significance of this result for the security of protocol E1, note that it also implies that Alice’s output bit must look somewhat random to the eavesdropper, even when provided with Alice and Bob’s inputs. To see this, observe that any such eavesdropper could be lumped together with Bob to derive a successful strategy in the guessing game described above.

The proof is in the form of a reduction to the following trivial guessing game. There are two cooperating players, Alice and Bob. At the start of the game, Bob receives a single bit $y \in \{0, 1\}$ chosen uniformly at random. The players are then allowed to perform arbitrary computations, but are not allowed to communicate. At the end of the game, Alice outputs a bit a , and the players win if $a = y$. Clearly, any strategy with success probability that deviates from $\frac{1}{2}$ indicates a violation of the no-signaling assumption between Alice and Bob. The idea behind the use of guessing games is to show that any undesirable outcome in the cryptographic protocol—such as a successful strategy for the Eavesdropper—can be bootstrapped into a successful strategy for Alice and Bob in the trivial guessing game, thereby contradicting the no-signaling assumption.

GUESSING LEMMA. Suppose that Alice and Bob succeed in the augmented CHSH game with probability at least $\frac{2}{3}\omega + \frac{1}{3} - \eta$, where $\omega = \cos^2 \frac{\pi}{8}$. Then the maximum probability with which Bob can successfully guess Alice’s output on input $x = 2$ is at most $1 - C(\eta_0 - \eta)$, where C and η_0 are universal constants.

Although in general optimization over quantum strategies is intractable, good upper bounds can often be obtained by considering convex relaxations of the set of valid quantum strategies for the players (using e.g. semidefinite programming). Here the game is sufficiently simple that an intuitive argument can be given which already provides a meaningful bound.

PROOF. For simplicity, we start by showing that when $\eta = 0$, then Bob cannot guess Alice’s output with probability 1. A more careful accounting of the parameters then provides the claimed bound. Note that $\frac{2}{3}\omega + \frac{1}{3}$ is the maximum probability with which Alice and Bob can succeed in the augmented CHSH game—for this to happen they must win the CHSH game with optimal probability ω , and on inputs $x = 2$ and $y = 0$, the outputs a and b must be equal. We assume for contradiction that Bob can guess Alice’s output (on input $x = 2$) with probability 1.

We imagine repeated sequential execution of the devices, with the same fixed inputs x and y , for some number k of executions. Let \vec{b}_y , \vec{c}_y denote Bob’s outputs, and \vec{a}_x denote Alice’s outputs, over these k executions. Then the condition on the augmented CHSH game implies that the players’ outputs should match when the inputs are $(x, y) = (2, 0)$: $\vec{a}_2 = \vec{b}_0$. Also by the assumption of successful guessing, on inputs $(2, y)$ for any y the players’ outputs satisfy $\vec{a}_2 = \vec{c}_y$. It follows that $\vec{b}_0 = \vec{c}_0$. The fact that Alice and Bob win the CHSH game with probability ω implies that for any $x \in \{0, 1\}$, the relative Hamming distance $d_H(\vec{a}_x, \vec{b}_0) \approx (1 - \omega)$ (by the Chernoff bound this is sharply concentrated as k grow). It follows that $d_H(\vec{a}_x, \vec{c}_0) \approx (1 - \omega)$.

Now suppose Alice knew Bob’s output \vec{c}_y (we will justify this assumption later). We claim that this gives Alice an advantage in guessing Bob’s input y thus providing a contradiction with the elementary guessing game described at the start of the section. Alice’s strategy is as follows: if given her input x and output \vec{a}_x , she guesses $y = 0$ if $d_H(\vec{a}_x, \vec{c}_y) < (1 - \omega) + \epsilon$, and $y = 1$ otherwise, where ϵ is an arbitrarily small constant (depending on k).

First note that if $y = 0$ then she succeeds with probability close to 1. This is because as shown above, since $\vec{b}_0 = \vec{c}_0$, for any $x \in \{0, 1\}$, $d_H(\vec{a}_x, \vec{c}_0) \approx (1 - \omega)$. On the other hand, when $y = 1$, by the CHSH game condition it should be the case that $d_H(\vec{a}_0, \vec{b}_1) \approx 1 - \omega$ and $d_H(\vec{a}_1, \vec{b}_1) \approx \omega$. By the triangle inequality, $d_H(\vec{a}_0, \vec{a}_1) \geq 2\omega - 1 \approx .7$. It follows that \vec{a}_0 and \vec{a}_1 cannot both be close to any fixed \vec{c}_1 : i.e. both conditions $d_H(\vec{a}_0, \vec{c}_1) < (1 - \omega) + \epsilon$ and $d_H(\vec{a}_1, \vec{c}_1) < (1 - \omega) + \epsilon$ cannot be simultaneously satisfied. This means that in the case $y = 1$, Alice must succeed with probability about 1/2, and therefore overall Alice can guess Bob’s output with probability close to 3/4. Contradiction.

To justify the assumption that Alice knows Bob’s output \vec{c}_y , we consider the following experiment: Alice and Bob execute D_A and D_B in chunks of k executions, repeatedly. In each chunk, Alice chooses a uniformly random input $x \in \{0, 1\}$. Bob always chooses the same secret input y . Bob select a value \vec{c} uniformly at random and post-selects on those chunks where $\vec{c}_y = \vec{c}$. He communicates the indices of those chunks to Alice. The information this leaks about y is very small, since the marginal distribution of \vec{c}_y must match the marginal distribution \vec{a}_2 obtained when D_A is provided input $x = 2$, which does not depend on y ; as a consequence the indices of the

chunks sent by Bob to Alice are close to uniformly distributed, irrespective of y .

This concludes the proof for the case of $\eta = 0$. It is not difficult to reproduce the same proof to make the reasoning more quantitative in the more general case that $\eta > 0$. We leave this as an exercise to the reader. \square

Before continuing with the multi-round analysis, we note that the guessing lemma already gives an “in principle” proof of security for quantum key distribution, following an argument due to Barrett et al.⁴ (though they used a different, more complex Bell inequality to achieve a security parameter that scales inverse proportionally to the number of rounds). Consider the following simplified variant of Protocol E1. The users instruct their devices to sequentially play $\Theta(N)$ instances of the augmented CHSH game. Alice chooses a random time at which to stop the games, and announces it (publicly) to Bob. The users publicly exchange their inputs. Once this has been done, they determine the last round, prior to the stopping time, where they used the input pair $(2, 0)$, and should thus have matching outputs; call this the key round. They exchange their outputs in all rounds prior to the key round and estimate the (augmented) CHSH game correlations. Provided these correlations come close enough to the CHSH game bound they use their output in the key round for the final key. Based on a simple martingale argument it is not hard to show that, conditioned on the devices producing outputs that satisfy the CHSH game correlations in all rounds prior to the key round, the devices in the key round have a high probability of satisfying the CHSH game condition as well (even if it is not checked). Using the guessing lemma it must be that, in that round, the eavesdropper has a bounded probability of guessing the users’ shared bit.

Note that the above argument provides a complete proof of security (for a single bit of key), without having to resort to the quantum formalism at all: the only assumption needed for security is the no-signaling principle. Going beyond a single round, however, will require us to assume in addition that the devices’ behavior can be modeled using quantum mechanics.

6. EXTRACTING A SECURE KEY

We proceed to analyze the complete protocol^d. Our goal is to provide a reduction: we aim to show that any global strategy for the Eavesdropper that yields even partial information about the users’ final shared key implies an attack on a single round of the protocol of the form described, and ruled out, in the previous section. As mentioned earlier, analyzing multiple rounds of the protocol runs into fundamental obstacles having to do with the quantum nature of the Eavesdropper and her attack. We will get around these obstacles by appealing to powerful results from the theory of randomness extractors to directly deal with correlations between the classical information generated by Alice, Bob and the Eavesdropper.

^d The appropriate notion of entropy is the quantum conditional minentropy, which has an operational interpretation: it is simply the maximum probability with which the Eavesdropper can successfully make a correct guess for the complete string K , by performing an arbitrary measurement on her quantum side information.

The first step is to formalize exactly what constitutes a strategy for the Eavesdropper. Recall that protocol E1 includes an ultimate post-processing step called *privacy amplification*. This task converts the raw key K shared by Alice and Bob (which they derived from K_A and K_B by performing information reconciliation) into a shorter string Z . Privacy amplification guarantees that as long as K contains enough entropy from the viewpoint of the Eavesdropper, Z is indistinguishable from uniform, again from the Eavesdropper’s point of view. Privacy amplification is achieved by using a strong (quantum-proof) randomness extractor:^{12,32} a procedure Ext which takes two strings of bits as input, the source X (whose role will be played by K) and a short uniformly random seed Y (that will be generated using private randomness, and shared publicly), and combines them to produce an output $Z = \text{Ext}(X, Y)$. Think of the seed as a short tag used to select a hash function from a publicly specified family; the output is given by the evaluation of the chosen hash function on the source. The security condition for a strong extractor guarantees that no adversary, *even given access to the seed*, is able to distinguish the output of the extractor from a uniformly random string of the same length. This guarantee will be met, provided the source contains enough entropy from the viewpoint of the adversary.

The major challenge thus remains to establish that the source, K , contains enough entropy from the viewpoint of the Eavesdropper. Our approach to showing this is based on (the quantum generalization of) a proof technique from the theory of pseudo-randomness called the “reconstruction paradigm,” originally introduced by Trevisan³³ towards the analysis of a class of *strong randomness extractors*. Roughly, this paradigm allows us to make a stronger statement than the generic strong extractor reduction sketched above: it allows us to show directly that if the key generated by the protocol (after privacy amplification) is distinguishable from random by the eavesdropper, then there is an (efficient) classical procedure to reconstruct Alice’s raw key with non-negligible probability. This allows us to use classical information-theoretic tools to complete a reduction to the guessing game.

6.1. Global attacks

Before delving into more details about the reconstruction paradigm, we first review some of the difficulties to be overcome by the analysis.

Recall that the main difference introduced in protocol E1 compared to Ekert’s protocol is the presence of only a small fraction of test rounds. The following simple attack demonstrates that the restriction is necessary for security. Note that Alice’s device D_A is able to distinguish key rounds from test rounds, since it is given a special input, 2, for the former type of round. Suppose the eavesdropper programs D_A in a way that the device systematically outputs any bit on which its input is $x = 2$ twice: in the key round, as required, as well as in the round that immediately follows. Assuming there are comparatively few key rounds, it is likely that the second round will be a test round. The devices will probably fail the CHSH game condition in that round, but provided there are sufficiently many test rounds the effect on their average success probability will remain small. But outputs in test rounds are publicly exchanged by the users: the complete raw key is leaked to the Eavesdropper! This

attack illustrates the main difficulty faced in our scenario: the users' devices may have memory, and behave adaptively in each round, leveraging the users' public communication to covertly leak information about the final key.

Even setting the issue of adaptive devices aside, there is a second important difficulty. Assume the users had a way to enforce that their devices perform independent and identical operations in each round. It may still be the case that the Eavesdropper, on which no control is possible, benefits from a global measurement on all the side information she has collected throughout the protocol in order to implement her final attack on the key. Such an attack should be ruled out, even if its success probability is tiny, as long as it is higher than the security parameter for the protocol. So assume the Eavesdropper has performed a successful attack. Here begins the difficulty: conditioning on success of an unlikely event, which involves a measurement on the quantum side information and the public information leaked throughout the protocol, introduces correlations between the classical data observed by the users (including the devices' inputs and outputs in all rounds) that need to be taken care of by the security proof.

The confounding factor here is that Bayes' rule for conditional probabilities does not work in the presence of quantum side information. Informally, if $P_g(K|E)$ is the maximum probability with which the Eavesdropper may guess the string K , given her quantum side information E , then it is not the case that $P_g(K|E) = P_g(K_N|K_{N-1}, \dots, K_1, E) \dots P_g(K_1|E)$. In fact, it is possible for the quantities on the right-hand side to all be very close to 1, while the quantity on the left-hand side is very small, the reason being that guessing measurements for K_1, K_2, \dots, K_N , cannot necessarily be "stitched together" into a guessing measurement for the complete string.

In order to side-step the issue we treat the combination of the joint quantum state of the devices and the eavesdropper, and the possible effects of the global measurement, as a black box, and focus on the classical information processed by Alice and Bob and the ways it which it can correlate with classical information generated by the Eavesdropper. The main tool in the analysis is the "reconstruction paradigm" mentioned above, which we discuss next.

6.2. Security and the reconstruction paradigm

Let us first recall the main ideas behind the (classical) reconstruction paradigm.^{33,e} Consider a strong extractor Ext which takes two strings of bits as input, the source X and the seed Y , and combines them to produce an output $Z = \text{Ext}(X, Y)$. Proceeding by contradiction, assume the adversary has the ability to distinguish the output of the extractor from uniform, with non-negligible advantage ε . Observe the weakness of this starting point: the adversary's advantage could come from any kind of information; for example, she is able to predict, with small advantage ε , the parity of a small subset of the bits of Z , the location of which itself depends on other bits of Z , etc. The reconstruction paradigm uses properties of a specific construction of Ext to show that from any such adversary it is possible to construct (explicitly, and this

will be important for us) the following stronger adversary. The stronger adversary has the ability, given as side information the information available to the original adversary (partial information about X and the seed Y) as well as a few additional "advice bits" about X , to produce a "reconstructed" guess for the complete string X that is correct with probability of the order of $(\varepsilon/m)^2$, where m is the length of Z .

Our security proof relies on an adaptation of the reconstruction paradigm to the case of quantum side information. This allows us to argue that any eavesdropper who is able to distinguish the final key Z generated by the users from a uniformly random string can be "bootstrapped" into a stronger eavesdropper who, given a few additional bits of advice, is able to guess the complete string of outputs K_A generated by Alice's device in the protocol with non-trivial probability. Note that this probability is of the same order as the security parameter ε . To see that it is a strong bound, note that it is much larger than the easy bound of inverse exponential in the extracted key length, which follows (using any strong extractor) from the correspondence between quantum conditional min-entropy and guessing probability. Formally introducing the quantum reconstruction paradigm would require notation that is beyond the scope of this article. For the expert, we mention that the main tool used in the analysis is the "pretty good measurement" of Hausladen and Wootters,¹⁷ which allows us to gain a handle on the Eavesdropper's global measurement.

6.3. Reduction to the guessing game

Recall that the raw key K_A is formed from the bits $K_A = a_1, \dots, a_{|\mathcal{C}|}$ generated by Alice's device D_A in the key rounds $\mathcal{C} \subseteq \{1, \dots, N\}$. Suppose for contradiction that, when an extractor Ext_c built according to the specifications of the reconstruction paradigm is applied to K_A , with a uniform choice of seed, the output Z is *not* indistinguishable from uniform: there is an attack for the adversary which uses all the quantum side information E available to the Eavesdropper at the end of protocol E1 to distinguish Z from uniform with some advantage ε . Our first step is to apply the quantum reconstruction paradigm to place ourselves in a stronger position: as argued in the previous section, it follows that there exists a "bootstrapped" adversary whom, using a combination of the side information E and a small number of additional advice bits, is able to produce a guess for the string K_A that is correct with probability of order $(\varepsilon/m)^2$, where m is the length of Z .

The second step in the analysis is to rule out such attacks. We achieve this by performing a reduction to the guessing game considered in Section 5. In order to present the reduction, it is convenient to abstract the present scenario as the following multi-round form of the guessing game. Alice gets an N -trit input x , and Bob an N -bit input y , chosen uniformly at random. They use their respective devices, D_A and D_B , sequentially to generate N -bit outputs a and b . A third player, Eve, is given x, y , and a small number of arbitrary bits of advice about Alice and Bob's outputs. (This includes the outputs in the test rounds and the bits of advice required for the reconstruction procedure.) Alice and Bob's devices are assumed to satisfy the CHSH game correlations on average, when tested on a randomly chosen fraction γ of the rounds, with non-negligible probability of order ε over an execution of the protocol. The goal is to show

^e The parameters we give here are inaccurate, and are only meant to give an indication of the procedure.

that Eve is unable to produce a correct guess for the sub-string a_C , where C contains those rounds in which $(x, y) = (2, 0)$, with non-negligible probability.

To show this we perform a reduction to the single-round guessing game. We seek to identify a round $i_0 \in \{1, \dots, N\}$ such that the following holds. Let D_{A,i_0} and D_{B,i_0} denote the state of Alice and Bob's devices at the beginning of round i_0 , conditioned on all past events. This includes fixing values for all prior input and output bits, and making those values available to the three players. It should then be the case that (i) the devices D_{A,i_0} and D_{B,i_0} have a large success probability in the augmented CHSH game, when provided uniformly random inputs $(x, y) \in \{0, 1, 2\} \times \{0, 1\}$, and (ii) Eve has a strategy which allows her to produce a correct guess for the output a obtained by D_{A,i_0} , when given the input $x = 2$, with high probability. If both conditions can be shown to hold simultaneously, the single-round guessing lemma, Lemma 1, will give a contraction.

As explained earlier, the main difficulty in completing the reduction is that the adversary's attack in the original, multi-round protocol is based on a global measurement on her side information, which includes the classical information gleaned throughout the protocol, as well as her initial quantum state. The application of the quantum reconstruction paradigm dispenses with the need to deal with the adversary's quantum state at the expense of providing her with a small number of advice bits. We separate this information in two parts: first, the users' input strings (x, y) . Second, the user's outputs (a_B, b_B) in test rounds, and the advice bits $g(a_C)$. This second part of the information can be handled using a standard trick: instead of waiting for the bits to be available, we modify the adversary into one which makes a uniformly random choice for them. Using that the number of rounds used as checks is small compared to $|C|$, her success probability remains non-negligible.

The other part of the side information, the inputs (x, y) , cannot be eliminated in the same way, as it contains too many bits. We can split the inputs in two parts: those to be chosen before round i_0 , and those chosen after. The inputs chosen before round i_0 will be made publicly available as part of the specification of the devices D_{A,i_0} and D_{B,i_0} , so there is no need to worry about them. Regarding the inputs to be chosen after round i_0 , it is enough to argue that a "good" choice of inputs exist: indeed, the device's output in round i_0 cannot depend on inputs provided to the device after round i_0 ; here the sequential nature of the protocol is used in an important way.

Now that we have a single, fixed measurement for the adversary, with hard-wired values for (x, y) and the advice bits, we are finally in a position to apply the (classical!) Bayes' rule in order to identify the round i_0 satisfying conditions (i) and (ii). This provides the following straightforward implication:

$$\begin{aligned} & \Pr(\text{Eve guesses } a_1 \dots a_{|C|} \text{ correctly}) \geq \varepsilon \\ & \Rightarrow \exists i_0, \Pr(\text{Eve guesses } a_{i_0} \mid \text{guesses for } a_1, \dots, a_{i_0-1} \text{ were correct}) \geq 1 - \delta, \end{aligned} \quad (1)$$

for some small constant $\delta > 0$.

The conditioning implied by Bayes' rule, however, presents a last difficulty: it may introduce correlations between the state of the devices at the beginning of the i_0 -th round, and the inputs that the devices "expect" in that round. Indeed, the conditioning

is on the event that the adversary's guesses in rounds prior to i_0 are correct. The guesses in turn depend on the choice of inputs (x, y) that were "hard-wired" into the adversary as part of the reduction described above. Thus correctness of the adversary's guesses can bias the distribution of inputs in round i_0 , which is no longer uniform and may be jointly correlated with the state of the devices D_{A,i_0} and D_{B,i_0} . For example, it could be that the inputs in a fixed round i_0 are forced to a specific choice such as $(0, 0)$, making the guarantee (i) all but useless (if the inputs are fixed, it is easy to win the CHSH game). This difficulty is similar to one encountered in the analysis of the parallel repetition of multiplayer games. The standard approach for this problem was introduced in work of Raz²⁶ showing a parallel repetition theorem for the value of classical two-player games. The main idea consists in arguing that conditioning on an event with large enough probability cannot introduce strong correlations in all, or even most, coordinates of a distribution that is initially in product form. At a technical level, the analysis uses the chain rule for mutual information and Pinsker's inequality.

In our scenario a very similar approach can be employed to show that conditioning on the adversary's success, provided success happens with large enough probability, cannot bias the users' input distribution in most rounds by too much. This uses that initially (before the conditioning) the distribution is a product distribution, sampled independently in each of the large number of rounds of the protocol.

Unfortunately this statement is not sufficient for our purposes. We need to show, not only that the inputs to the devices in the round i_0 are close to uniformly random, but also that the post-measurement state of the devices (conditioned on the same success event) is not correlated to the users' choice of input. To show that this cannot happen we expand on Raz' technique by introducing a coding argument to deal with the quantum correlations. Assume for contradiction that the devices' state has a non-trivial correlation with their input (after a successful guess of the eavesdropper), and that this holds in most rounds. We show that such devices, including the eavesdropper's strategy, could be used to transmit classical information from Bob and the eavesdropper to Alice at a more efficient rate than is allowed by standard arguments from quantum information theory, thereby reaching a contradiction. For a further discussion of this step we refer to previous work of the authors.³⁵

7. OUTLOOK

From their origins in the EPR thought experiment, through their formulation in Bell's work and the CHSH test, to their use in device-independent cryptography, the non-local correlations of quantum mechanics have gone from undesirable paradox to operationally desirable certificates, not only of quantumness but also of randomness and privacy.^f

Our proof of security inscribes itself in a long sequence of works demonstrating progressively stronger uses of quantum

^f Note that the conditioning is performed jointly on an event involving Alice and Bob (the CHSH violation observed in previous rounds being sufficiently large) on the one hand, and Bob and Eve (Eve's guess being correct) on the other, so it can certainly introduce correlations across the whole input distribution.

correlations. Although the relevance of non-local correlations for cryptography was already pointed out by Ekert, the first concrete results considered the related task of randomness amplification, in which a single user, Alice, wishes to certify that two devices in her possession generate random bits, while using as few possible seed bits of randomness to test the device.

A protocol for randomness certification was first proposed in Colbeck,¹⁰ and the task was experimentally demonstrated in 2012.²⁵ Some of the ideas present in this paper were first formulated in the analysis of a protocol for exponential randomness expansion we introduced in previous work of the authors;³⁴ subsequently it was shown that even unbounded expansion is possible.¹¹

Subsequent to our work on quantum key distribution a second proof of security for DIQKD was put forward in Miller and Shi.²³ Neither proof achieves practical parameters, a gap recently filled by a third proof³ which establish security of a protocol with essentially optimal noise tolerance and key rate. Although the three proofs inevitably bear some similarity, they seem to rely on essentially different arguments; it remains a challenge to find a unifying framework for device-independence that would combine their respective advantages.

Recent progress in experiments^{15,18,30} place the implementation of protocols for entanglement-based quantum key distribution just around the corner. One may even start to see emerging the possibility for entanglement generation not only between single-qubit devices, as are needed for QKD, but small multi-qubit “quantum computers in the cloud” such as the one recently demonstrated by IBM. Computing devices sharing quantum entanglement may be able to implement tasks beyond quantum key distribution, from simple protocols such as, for example, quantum secret sharing or entanglement swapping, to the complex task of verifiable delegated computation. Although this remains a distant prospect, it is our hope that the techniques developed in this work will find far broader applicability in the future of quantum cryptography!

Acknowledgments

We thank the *Communications* editors and Gilles Brassard for useful feedback on an earlier version of this article.

Umesh Vazirani is supported by NSF Grant CCF-1410022, MURI Grant FA9550-18-1-0161 and a Vannevar Bush Faculty Fellowship. Thomas Vidick is supported by AFOSR YIP award number FA9550-16-1-0495, MURI Grant FA9550-18-1-0161, and a CIFAR Azrieli Global Scholar award. □

References

- Acín, A., Brunner, N., Gisin, N., Massar, S., Pironio, S., Scarani, V. Device-independent security of quantum cryptography against collective attacks. *Phys. Rev. Lett.* 98, 230501 (2007).
- Acín, A., Gisin, N., Masanes, L. From Bell's theorem to secure quantum key distribution. *Phys. Rev. Lett.* 97, 120405 (2006).
- Arnon-Friedman, R., Renner, R., Vidick, T. Simple and tight device-independent security proofs. *arXiv preprint arXiv:1607.01797* (2016).
- Barrett, J., Colbeck, R., Kent, A. Unconditionally secure device-
- independent quantum key distribution with only two devices. *Technical report arXiv:1209.0435* (2012).
- Barrett, J., Hardy, L., Kent, A. No signaling and quantum key distribution. *Phys. Rev. Lett.* 95, 010503 (2005).
- Bell, J.S. On the Einstein-Podolsky-Rosen paradox. *Physics* 1 (1964), 195–200.
- Bennett, C., Brassard, G. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of the International Conference on Computers, Systems, and Signal Processing* (1984), 175–179.

- Brassard, G., Lütkenhaus, N., Mor, T., Sanders, B.C. Limitations on practical quantum cryptography. *Phys. Rev. Lett.* 85 (Aug 2000), 1330–1333.
- Clauser, J.F., Horne, M.A., Shimony, A., Holt, R.A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.* 23 (1969), 880–884.
- Colbeck, R. *Quantum and Relativistic Protocols for Secure Multi-Party Computation*. PhD thesis, Trinity College, University of Cambridge, Cambridge, UK, Nov. 2006.
- Coudron, M., Yuen, H. Infinite randomness expansion with a constant number of devices. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing* (ACM, 2014), 427–436.
- De, A., Portmann, C., Vidick, T., Renner, R. Trevisan's extractor in the presence of quantum side information. *SIAM J. Comp.* 41, 4 (2012), 915–940.
- Einstein, A., Podolsky, P., Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* 47 (1935), 777–780.
- Ekert, A.K. Quantum cryptography based on Bell's theorem. *Phys. Rev. Lett.* 67 (1991), 661–663.
- Giustina, M., Versteegh, M.A., Wengerowsky, S., Handsteiner, J., Hochrainer, A., Phelan, K., Steinlechner, F., Kofler, J., Larsson, J.-Å., Abellán, C., et al. Significant-loophole-free test of Bell's theorem with entangled photons. *Phys. Rev. Lett.* 115, 25 (2015), 250401.
- Goldwasser, S., Micali, S. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *Proceedings of the fourteenth annual ACM symposium on Theory of computing* (ACM, 1982), 365–377.
- Hausladen, P., Wootters, W.K. A “pretty good” measurement for distinguishing quantum states. *J. Mod. Opt.* 41, 12 (1994), 2385–2390.
- Hensen, B., Bernien, H., Dréau, A.E., Reiserer, A., Kalb, N., Blok, M.S., Ruitenberg, J., Vermeulen, R.F., Schouten, R.N., Abellán, C., et al. Loophole-free bell inequality violation using electron spins separated by 1.3 kilometers. *Nature* 526, 7575 (2015), 682–686.
- Kalai, Y.T., Raz, R., Rothblum, R.D. How to delegate computations: the power of no-signaling proofs. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing* (ACM, 2014), 485–494.
- Lydersen, L., Wiechers, C., Wittmann, C., Elser, D., Skaar, J., Makarov, V. Hacking commercial quantum cryptography systems by tailored bright illumination. *Nat. Photonics* 4, 10 (2010), 686–689.
- Mayers, D. Unconditional security in quantum cryptography. *J. ACM* 48, 3 (May 2001), 351–406.
- Mayers, D., Yao, A. Quantum cryptography with imperfect apparatus. In *Proceedings of the 39th Annual Symposium on Foundations of Computer Science, FOCS '98* (IEEE Computer Society, 1998), Washington, DC, USA, 503.
- Miller, C.A., Shi, Y. Robust protocols for securely expanding randomness and distributing keys using untrusted quantum devices. *J. ACM* 63, 4 (2016), 33.
- Pironio, S., Acín, A., Brunner, N., Gisin, N., Massar, S., Scarani, V. Device-independent quantum key distribution secure against collective attacks. *New J. Phys.* 11, 4 (2009), 045021.
- Pironio, S., Acín, A., Massar, S., De La Giorday, A.B., Matsukevich, D.N., Maunz, P., Olmschenk, S., Hayes, D., Luo, L., Manning, T.A., et al. Random numbers certified by Bell's theorem. *Nature* 464, 7291 (2010), 10.
- Raz, R. A parallel repetition theorem. *SIAM J. Comput.* 27 (1998), 763–803.
- Reichardt, B.W., Unger, F., Vazirani, U. Classical command of quantum systems. *Nature* 496, 7446 (2013), 456.
- Rivest, R.L., Shamir, A., Adleman, L. A method for obtaining digital signatures and public-key cryptosystems. *Comm. ACM* 21, 2 (1978), 120–126.
- Scarani, V., Gisin, N., Brunner, N., Masanes, L., Pino, S., Acín, A. Secrecy extraction from no-signaling correlations. *Phys. Rev. A* 74, (Oct 2006), 042339.
- Shalm, L.K., Meyer-Scott, E., Christensen, B.G., Bierhorst, P., Wayne, M.A., Stevens, M.J., Gerrits, T., Glancy, S., Hamel, D.R., Allman, M.S., et al. Strong loophole-free test of local realism. *Phys. Rev. Lett.* 115, 25 (2015), 250402.
- Shor, P.W., Preskill, J. Simple proof of security of the BB84 quantum key distribution protocol. *Phys. Rev. Lett.* 85 (Jul 2000), 441–444.
- Tomamichel, M., Schaffner, C., Smith, A., Renner, R. Leftover hashing against quantum side information. *IEEE Transactions on Information Theory* 57, 8 (2011), 5524–5535.
- Trevisan, L. Extractors and pseudorandom generators. *J. ACM* 48 (July 2001), 860–879.
- Vazirani, U., Vidick, T. Certifiable quantum dice: Or, true random number generation secure against quantum adversaries. In *Proceedings of the 44th symposium on Theory of Computing, STOC '12* (ACM, 2012), 61–76. Also available as arXiv:1111.6054.
- Vazirani, U., Vidick, T. Fully device-independent quantum key distribution. *Phys. Rev. Lett.* 113, 14 (2014), 140501.
- Yin, J., Cao, Y., Li, Y.-H., Liao, S.-K., Zhang, L., Ren, J.-G., Cai, W.-Q., Liu, W.-Y., Li, B., Dai, H., et al. Satellite-based entanglement distribution over 1200 kilometers. *Science* 356, 6343 (2017), 1140–1144.

Umesh Vazirani (vazirani@eecs.berkeley.edu), Department of Computer Science, UC Berkeley, Berkeley, California, USA.

Thomas Vidick (vidick@cms.caltech.edu), Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California, USA.

[CONTINUED FROM P. 144] five kilometers, thus one more kilometer than the first player.

Question 1. By playing differently, beginning with the first move, could the first player acquire the rights to more of the line segment than the second player?

Solution to question 1. Yes. If the first player takes kilometers 2 to 3, then the second player could take kilometers 0 to 2, but then the first player would take kilometers 3 to 5. The first player would thus get three of the five kilometers.

Question 2. Is there a minimal amount by which one player can win regardless of the length L of the segment?

Answer. Yes. The first player can win by at least one kilometer every time by going in the middle, meaning the halfway point of the first player's kilometer is at position $L/2$. After that, the first player would mirror the second player's moves. So if the second player takes x to $x+2$ to the left of

the middle kilometer, then the first player takes $(x + L/2)$ to $(x + 2 + L/2)$ to the right of the middle. The net effect is the first player can always guarantee to capture at least as much territory as the second player on the two sides of that middle kilometer. The first player wins by at least the kilometer of the first move.

Upstart 1. Characterize situations in which the first player can guarantee to win by more than one kilometer or prove it cannot be done.

Upstart 2. Suppose the line segment is of length L , but there are now k players instead of two. The rules are a direct generalization of the original game; the first player may take one kilometer, the second player two, the third player three, ... the k^{th} player k , the first player then takes $k+1$... and so on, all without overlap. Is there some length L and some number of players k whereby a player other than the first player can guarantee to capture more of the line segment than anyone else?

Upstart 3. Suppose the government leased out vertical cross-sectional squares belowground. Each player would thus take squares, with the side length of each square increasing by one kilometer with each move. The first player takes one kilometer squared. The second player then gets two kilometers squared. The first player then gets three kilometers squared, and so on, again without overlap. Does either player have a winning strategy if the area available to lease could be an arbitrary rectangular cross-section belowground? How would this generalize to more players?

All are invited to submit their solutions to upstartpuzzles@cacm.acm.org; solutions to upstarts and discussion will be posted at <http://cs.nyu.edu/cs/faculty/shasha/papers/cacmpuzzles.html>

Dennis Shasha (dennishasha@yahoo.com) is a professor of computer science in the Computer Science Department of the Courant Institute at New York University, New York, USA, as well as the chronicler of his good friend the omnivorous Dr. Ecco.

Copyright held by author.



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
acmmEDIASales@acm.org

The National Academies of SCIENCES • ENGINEERING • MEDICINE

ARL Distinguished Postdoctoral Fellowships

The Army Research Laboratory (ARL) Distinguished Postdoctoral Fellowships provide opportunities to pursue independent research that supports the mission of ARL. The Fellow benefits by having the opportunity to work alongside some of the nation's best scientists and engineers. ARL benefits by the expected transfer of new science and technology that enhances the capabilities of the U.S. Army and the warfighter in times of both peace and war.

ARL invites exceptional young researchers to participate in this excitement as ARL Distinguished Postdoctoral Fellows. These Fellows must display extraordinary abilities in scientific research and show clear promise of becoming outstanding future leaders. Candidates are expected to have already successfully tackled a major scientific or engineering problem during their thesis work or to have provided a new approach or insight, evidenced by a recognized impact in their field. ARL offers five named Fellowships honoring distinguished researchers and work that has been performed at ARL. Three of these positions are open for the 2019 competition.

The ARL Distinguished Postdoctoral Fellowships are one-year appointments, renewable for up to three years based on performance. The annual stipend is \$100,000, and the award includes benefits and potential additional funding for the chosen proposal. Applicants must have completed all requirements for a Ph.D. or Sc.D. degree by the application deadline and may not be more than five years beyond their doctoral degree as of the application deadline. For more information and to apply, visit www.nas.edu/arl.

Online applications must be submitted by May 31, 2019 at 5 PM EST.





Dennis Shasha

DOI:10.1145/3314071

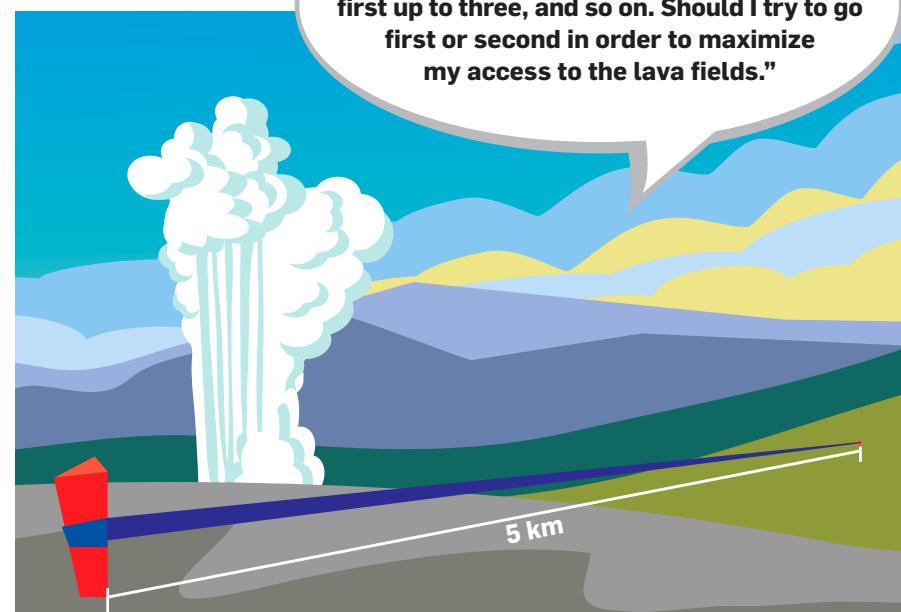
Upstart Puzzles Fighting for Lava

THE VAST UNDERGROUND lava fields in the western U.S. feature a photogenic geyser called Old Faithful. Eruptions send approximately 15,000 liters of steaming water 50 meters into the air approximately every hour. Unfortunately, what is underground is not nearly as appealing. If the lava fields erupted in a major way, they could cause ferocious firestorms that would destroy a large portion of the western U.S. and Canada and substantially cool the planet.

Now imagine a pair of tunnel-boring energy-extraction companies are competing to cool the lava, make some money, and provide carbon-free energy besides. The idea is to tunnel from a power plant outside the lava fields to near the lava, but not too close, to avoid accidental eruptions. A pipeline could in theory then take cool water from the power plant to the end of the tunnel where the water would be heated into steam and the steam would power the turbines of the power plant. The whole system could be designed to recycle the steam back into water in a closed loop.

In this scenario, the federal government, which owns the land, steps in to lease the energy rights, identifying cross-sections underground to which tunnels can be drilled. The government identifies those underground cross-sections based on their more-or-less linear segments aboveground, so leasing a segment would confer the right to tap the lava in the vertical cross-section below that segment.

To encourage participation in the project while achieving equity for both companies, government mathematicians design a game-style



protocol, whereby each company (player) takes turns to acquire non-overlapping sections of a full segment. The first player may take up

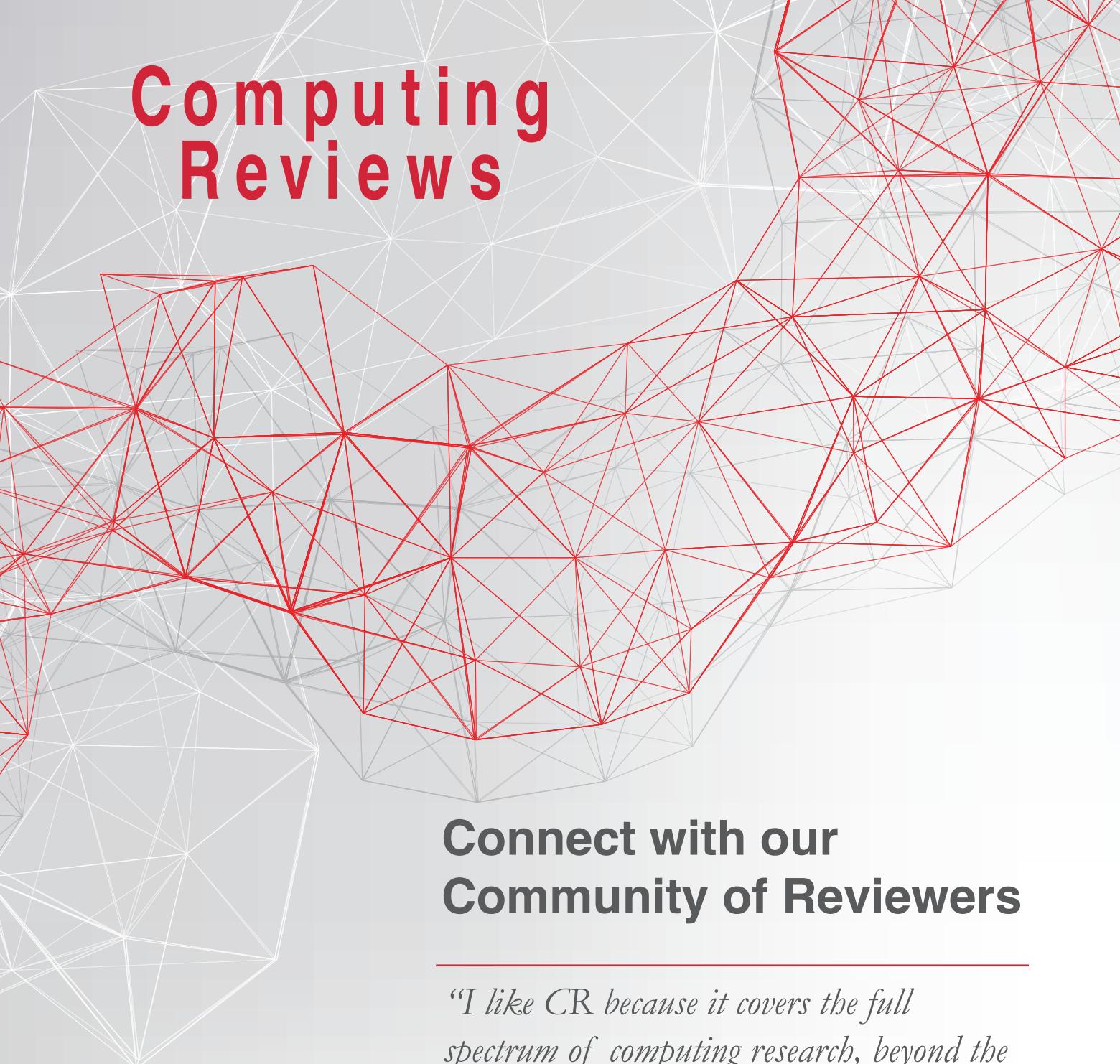
to one kilometer in the first turn, the second player then takes up to two kilometers, the first player then gets up to three kilometers, the second then gets up to four kilometers, and so on.

Imagine a pair of tunnel-boring energy-extraction companies are competing to cool the lava, make some money, and provide carbon-free energy besides.

Warm-up. Suppose the line segment is five kilometers long from a stake at kilometer 0 to a stake at kilometer 5. Suppose the first player takes between 0 and 1 kilometer. Which player would get more of the line segment, assuming each plays optimally?

Answer to warm-up. Player 2. Player 1 takes kilometer 0 to 1. Player 2 then takes kilometers 2 to 4. The first player then takes one of the two remaining kilometers—1 to 2 or 4 to 5—and the second player then takes the other remaining kilometer. The second player ends up with lease rights to three of the [CONTINUED ON P. 143]

Computing Reviews



Connect with our Community of Reviewers

"I like CR because it covers the full spectrum of computing research, beyond the comfort zone of one's specialty. I always look forward to the next Editor's Pick to get a new perspective."

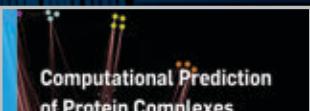
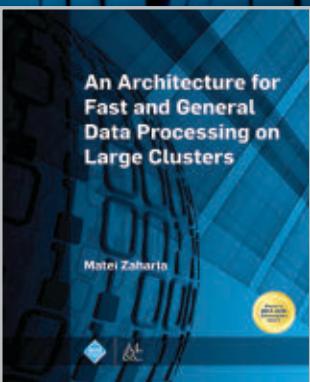
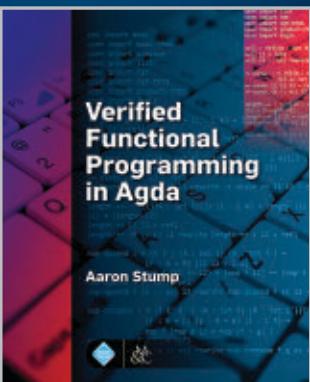
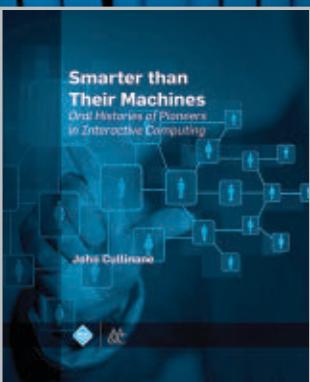
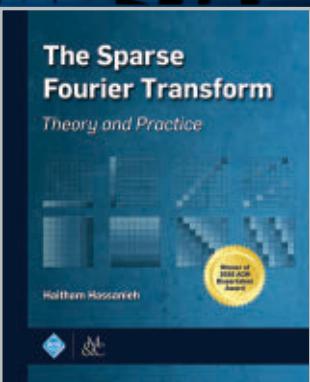
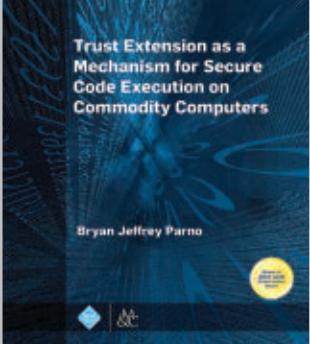
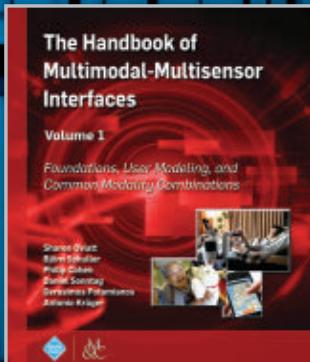
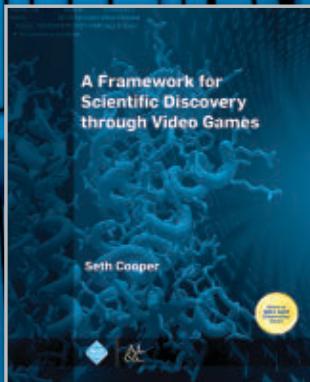
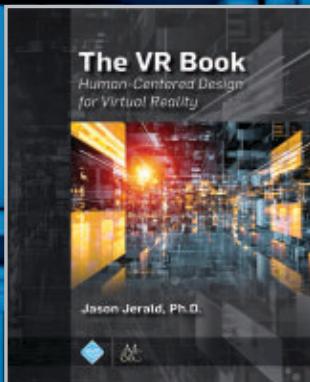
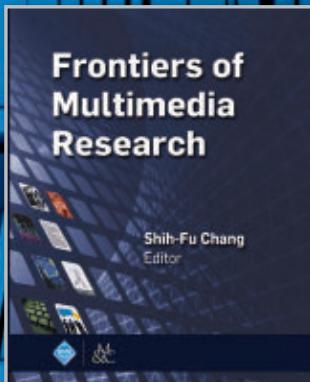
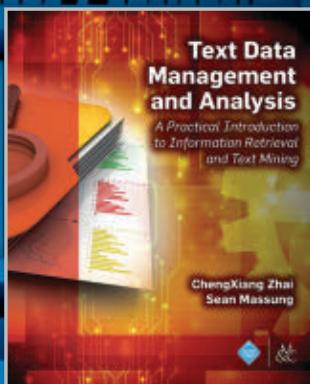
- Alessandro Berni



Association for
Computing Machinery

ThinkLoud

www.computingreviews.com



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

2 Penn Plaza, Suite 701, New York, NY 10121-0701, USA
 Phone: +1-212-626-0658 Email: acmbooks-info@acm.org