COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

11/2019 VOL.62 NO.11

Special Section on India Region

An Interview with Leonard Kleinrock

DeepXplore: Automated Whitebox Testing of Deep Learning Systems

When Drones Fly

Association for Computing Machinery

SIGGRAPH ASIA 2020 DAEGU

The 13th ACM SIGGRAPH Conference and Exhibition on Computer Graphics and Interactive Techniques in Asia

Conference 17 – 20 November 2020
Exhibition 18 – 20 November 2020
EXCO, Daegu, South Korea

Driving Diversity

SA2020.SIGGRAPH.ORG

#SIGGRAPHAsia | #SIGGRAPHAsia2020

Sponsored by

acm

Organized by

koelnmesse

TCS Research

Positions for PhDs: fresh and mid-career

Tata Consultancy Services (TCS) is a part of the Tata group, a multinational business group with interests across industry segments. TCS is an IT services, consulting and business solutions provider that has been partnering with the world's largest businesses in their transformation journeys for the last fifty years.

TCS operates on a global scale, with diverse talent base representing 149 nationalities, across 46 countries. TCS has been recognized as a Global Top Employer for the fourth consecutive year (2018-19) by the Top Employers Institute, and the Number One Top Employer in four regions – North America, Europe, Asia Pacific, and the Middle East.

TCS has been a pioneer in Software Research and has continued to systematically invest in Research over decades. TCS Researchers seek to make an impact and solve real-world problems for global Fortune 1000 companies. Towards this goal, TCS Research conducts applied industrial research in the following areas: Behavioral, Business, and Social Sciences, Computing Systems, Cybersecurity and Privacy, Data and Decision Sciences, Deep Learning and AI, Embedded Systems and Robotics, Foundations of Computing, Life Sciences, Media and Advertising, Physical Sciences, Software Systems and Services.

New York
Pittsburgh
Toronto
London
Delhi
Hyderabad
Mumbai
Chennai
Pune
Kolkata
Bangalore
CANADA
USA
UK
INDIA
TCS Research Labs
TCS Pace Ports™

TCS Research invites applicants for **Full time Research positions** at its labs across Indian Cities such as **Bangalore, Chennai, Delhi** (Noida and Gurugram), **Hyderabad, Kolkata, Mumbai,** and **Pune.**

We also invite applications for 2-3 year postdoctoral positions in our newly created TCS Pace Ports™ at **New York** (within the **Cornell Tech campus**), **Pittsburgh** (within the **Carnegie Mellon University campus**), **Toronto** (located near the **University of Toronto**), and **London** (located near the **Imperial College**).

Note: Positions located in NY and Pittsburgh will be in Fall 2019; Toronto and London would start in Fall 2020.

We seek researchers who will advance our capabilities in our core research areas, contribute to global thought leadership and create an intellectual foundation to address current and future business and technology opportunities.

Who can apply?
Applicants should have a PhD from a premier University/Institute related to the Research Areas mentioned above. Fresh PhDs as well as mid-career researchers are invited to apply.

How to apply:
Write to us at **careers.research@tcs.com** with a detailed CV highlighting your interest area of research, education and publications. Please mark **"Research Careers"** in the subject line.

For more details : **https://www.tcs.com/tcs-research**

TATA

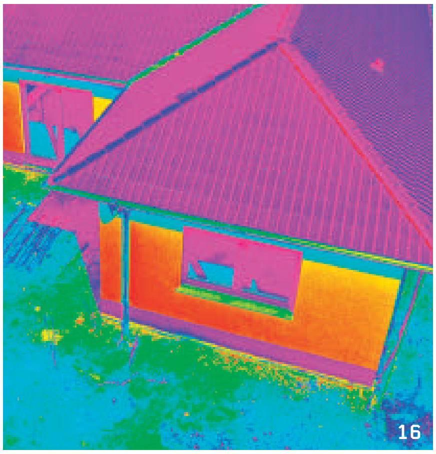
TATA CONSULTANCY SERVICES
Experience certainty.

COMMUNICATIONS OF THE ACM

Departments

Last Byte

News


16

Viewpoints

28

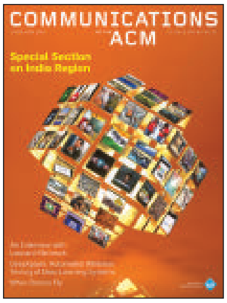
IMAGES BY: (L) RIKOBEST; (R) OLLYY

11/2019
VOL. 62 NO. 11

Special Section: India Region

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

About the Cover:

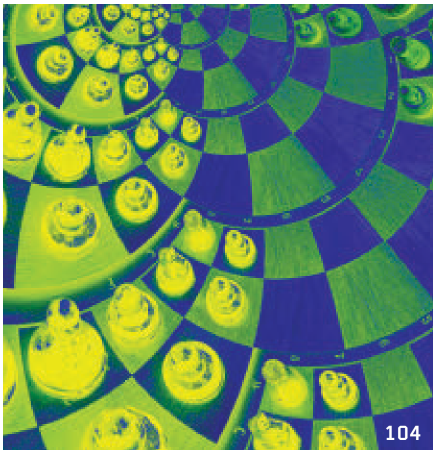

COMMUNICATIONS ACM
Special Section on India Region

The nations that make up the India region serve as a thriving nexus of technology innovations, research advances, and computing challenges. The articles in this month's special section, written by many of the area's leading lights, provide unique perspectives into the latest activities from this region. Cover illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: School photo courtesy of CSpathshala.org. Group photo courtesy of FSTTCS. Inmobi photo courtesy of Inmobi.com blog. Rivigo image courtesy of Rivigo.com. Pakistani women photo by Gary Yim/Shutterstock.com; Highway photo by SNEHIT/Shutterstock.com; Zomato photo by Jason Arora/Shutterstock.com; Wipro, Zoho, Flipkart photos by Piotr Swat/Shutterstock.com; Sign photo by Filip Jedraszak/Shutterstock.com; Ola car photo by Phuong D. Nguyen/Shutterstock.com; Aadhaar photos by Melting Spot/Shutterstock.com; Tata phone photo by Igor Golovniov/Shutterstock.com; Oyo photo by fotosunny/Shutterstock.com; Bagmane Tech Park photo by Noppasin Wongchum/Shutterstock.com; Traffic photo by sladkozaponi/Shutterstock.com; Office photo by CRS PHOTO/Shutterstock.com; Lenskart photo by Grzegorz Czapski/Shutterstock.com; Crowd photo by Dipak Shelare/Shutterstock.com; Tata building photo by Bilal Aliyar m/Shutterstock.com; Infosys building photo by Ajay Tvm/Shutterstock.com. Additional stock images from Shutterstock.com.

IMAGES BY: (L) KONSTANTIN FARAKTINOV; (R) MIKHAIL LEONOV, COLORIZED BY ANDRIJ BORYS ASSOCIATES

Practice



Articles' development led by acmqueue
queue.acm.org

Contributed Articles

Review Articles

Research Highlights

Watch the authors discuss their work in the exclusive *Communications* video.
https://cacm.acm.org/videos/deepxplore

acm
Association for Computing Machinery
Advancing Computing as a Science & Profession

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.

STAFF
DIRECTOR OF PUBLICATIONS
Scott E. Delman
cacm-publisher@cacm.acm.org

Executive Editor
Diane Crawford
Managing Editor
Thomas E. Lambert
Senior Editor
Andrew Rosenbloom
Senior Editor/News
Lawrence M. Fisher
Web Editor
David Roman
Editorial Assistant
Danbi Yu

Art Director
Andrij Borys
Associate Art Director
Margaret Gray
Assistant Art Director
Mia Angelica Balaquiot
Production Manager
Bernadette Shade
Intellectual Property Rights Coordinator
Barbara Ryan
Advertising Sales Account Manager
Ilia Rodriguez

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

CONTACT POINTS
Copyright permission
permissions@hq.acm.org
Calendar items
calendar@cacm.acm.org
Change of address
acmhelp@acm.org
Letters to the Editor
letters@cacm.acm.org

WEBSITE
http://cacm.acm.org

WEB BOARD
Chair
James Landay
Board Members
Marti Hearst; Jason I. Hong; Jeff Johnson; Wendy E. MacKay

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

ACM ADVERTISING DEPARTMENT
1601 Broadway, 10th Floor
New York, NY 10019-7434 USA
T (212) 626-0686
F (212) 869-0481

Advertising Sales Account Manager
Ilia Rodriguez
ilia.rodriguez@hq.acm.org

Media Kit acmmediasales@acm.org

Association for Computing Machinery (ACM)
1601 Broadway, 10th Floor
New York, NY 10019-7434 USA
T (212) 869-7440; F (212) 869-0481

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

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

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

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

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

Printed in the USA.

Association for
Computing Machinery

PLEASE RECYCLE THIS MAGAZINE

DOI:10.1145/3364216

Vinton G. Cerf

Hazards of the Information Superhighway

IN THE 1990S, U.S. Vice President Al Gore characterized the Internet as an "Information Superhighway." This metaphor has some utility as we try to understand emerging properties of the global Internet. More recently, an old friend, Judith Estrin, touted the importance of friction in the online environment. She had two things in mind, I believe. The first is that friction slows things down and sometimes that is exactly what is needed to give time to think about the content found on the Internet, especially in social media. Friction also keeps you on the road and not spinning off at every turn. As reports of the deliberate injection of misinformation and disinformation into the Internet continue to escalate, my attention has been drawn to efforts to counter this trend. I went back and re-read the May 2019 report about the Finnish response to information pollution,[a] which has garnered attention from other countries and organizations concerned about this phenomenon.

The Finnish response centers on critical thinking and teaching citizens of all ages to ask probing questions about information they gather whether online or offline. Propaganda is intended to steer the recipient's thinking into the directions intended by its source. Interestingly, the so-called weaponization of information need not be unidirectional. The disinformation campaigns allegedly conducted by Russia against the U.S., France, and the U.K., for example, were often designed to pit opposing groups against one another for the purpose of disrupting democracy. The propagandists were not interested in one group or another prevailing as much as they wanted to sow distrust of democratic institutions, disrupt rational and civil discourse, and generally increase domestic tensions among groups with potentially conflicting agendas.

It is tempting to think such mischief would be obvious to those exposed to these campaigns but we are human and being human we are subject to effects such as group think and confirmation bias. We grow comfortable with our beliefs and those of like-minded people, so much so that even in the face of clear evidence, we may be more likely to reject factual refutation of our positions than to change our minds and our positions. Indeed, there is some evidence that factual rebuttals may generate increased intolerance of views opposing our own, despite their factual basis.

The Finnish antidote is to train its citizens to think critically about what they see and hear; to ask questions about corroborating evidence; to explore and uncover the sources of controversial statements. That this takes real work is evident. Students report the effort is sometimes onerous. Nonetheless, it strikes me that such effort is an obligation derived from living in a democratic society. The price we pay for the freedom of access to information that we enjoy on the open Internet is the need for due diligence applied to the sources of information we rely upon.

Not surprisingly, brand can become a key indicator of quality of information if the branded source can be repeatedly validated. In the global Internet, there is a universe of sources and finding quality brands is made all the more difficult by the scale of the problem. Given the critical nature of the Internet's search engines as tools for discovery of World Wide Web content, it seems inescapable that the presentation of search results not only must be prioritized by some measure of quality but also that the ranking criteria must be clear and well understood. Transparency is our friend in this endeavor. This also applies to sources of information. Unvalidated sources or anonymous sources should be considered less trustworthy than strongly authenticated ones. This does not mean, however, that even a well-known source should be taken at face value. Just because a source is well identified does not mean it carries valid information.

Ultimately, this takes us back to critical thinking and the need for multiple reinforcing sources. There may be serious disagreements among legitimate sources of information as is often the case in scientific disputes. The solution to those problems almost always relies on obtaining more factual information and better interpretive theories. This should be the essence of democratic discourse and should not be replaced by fabricated information intended to mislead and derail genuine search for truth. C

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

[a] https://www.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/

CALL FOR PAPERS

2020 IEEE WORLD CONGRESS ON SERVICES

CLOUD/ICWS/SCC/SmartDataServices/DHAASS

July 6-11, 2020 BEIJING, CHINA

http://conferences.computer.org/services/2020/

IEEE SERVICES

IEEE

IEEE COMPUTER SOCIETY

TCSC

IBM Research

The 2020 IEEE World Congress on Services (SERVICES) will be held on July 6-11, 2020 in Beijing, China. The Congress is solely sponsored by the IEEE Computer Society under the auspice of the Technical Committee on Services Computing (TCSVC). The scope of the Congress will cover all aspects of services computing and applications, current or emerging. It covers various systems and networking research pertaining to cloud, edge and Internet-of-Things (IoT), as well as technologies for intelligent computing, learning, big data and blockchain applications, while addressing critical issues such as high performance, security, privacy, dependability, trustworthiness, and cost-effectiveness. The Congress will also include symposia and workshops supporting deep-dive discussions on emerging important topics, and complement the Congress program with industry and application presentations and panels. Authors are invited to prepare early and submit original papers to any of these conferences at www.easychair.org. All submitted manuscripts will be peer-reviewed by at least 3 reviewers. Accepted and presented papers will appear in the conference proceedings published by the IEEE Computer Society Press. The Congress will be organized with the following affiliated conferences and symposia:

IEEE International Conference on Cloud Computing (CLOUD): The flagship theme-topic conference for modeling, developing, publishing, monitoring, managing, delivering XaaS (everything as a service) in the context of various types of cloud environments.

IEEE International Conference on Web Services (ICWS): The flagship theme-topic conference for Web-based services, featuring Web services modeling, development, publishing, discovery, composition, testing, adaptation, and delivery, and Web services technologies as well as standards.

IEEE International Conference on Services Computing (SCC): The flagship theme-topic conference for services innovation lifecycle that includes enterprise and vertical services modeling, microservices based solution creation, services orchestration, services optimization, services management, services marketing, and business process integration and management.

IEEE International Conference on Smart Data Services (SmartDataServices): The flagship theme-topic conference for data driven applications and solutions under the as-a-service model, including analytic services, smart data foundation, big data services, blockchain, and data computing at the edge and in IoT systems.

Symposium on Digital Health as a Service (DHAASS): DHAASS represents an emerging and critical direction for SERVICES covering the application of digital health as a service in transforming health and social care. Key themes for 2020 include definitions and safe implementations of health/medical microservices (dubbed the Uber of digital health), crowd sensing/sourcing, microservice integration, health service economics, among others.

Key Dates: Early paper submission due December 2, 2019; Review comments for early-submission papers provided January 17, 2020; Normal paper submission due February 13, 2020; Final notification to authors provided April 6, 2020; Camera-ready manuscripts due April 20, 2020

Send inquiries to: ieeecs.services@gmail.com
See: http://conferences.computer.org/services/2020/ for more information.

DOI:10.1145/3365383

Moshe Y. Vardi

The Winner-Takes-All Tech Corporation

The five largest U.S. corporations—Alphabet, Amazon, Apple, Facebook, and Microsoft—are all tech companies with combined market capitalization of over four trillion dollars. Tech is often called "Big Tech" these days. Furthermore, a small number of corporations have come to dominate the IT industry, as within each industry segment one corporation often dominates.

The phenomenon whereby corporate dominance seems to be entrenched is often referred to as "winner takes all." In the context of tech, such a phenomenon can be partly explained by two "laws:" Metcalfe's Law asserts that the effect of a communications network is proportional to the square of the number of connected users. This makes Facebook, with over 1.5B daily users, dominant as a social network. Kai-Fu Lee's Virtuous Cycle asserts "More data begets more users and profit, which begets more usage and data." This explains, for example, the dominance of the Google search engine. Metcalfe's Law and the Virtuous Cycle make tech companies into natural monopolies, some claim.

As I argued earlier this year, we need laws and regulations, instead of an ethics outrage, to deal with undesired business models and conduct of tech corporations. What may have been a radical position less than a year ago has become a conventional wisdom now. There are several initiatives to regulate tech; the question now is how rather than if. The biggest regulatory issue on the table is how to deal with overly dominant corporations. In a 2018 book, *The Curse of Bigness: Antitrust in the New Gilded Age*, legal scholar Tim Wu argues the U.S. must enforce anti-trust laws against such corporations.

Public concerns about overly dominant corporations have been aggravated by what has become a dogma in the U.S. business community over the past generation, which is the Shareholder-Primacy Principle, asserting that shareholders should be assigned a priority relative to all other corporate stakeholders, such as employees, customers, and the like. According to this view, the goal of a corporation is just to generate profits, period! This approach, which has emerged in the 1970s and became dogmatic in the 1980s, has replaced the earlier approach of "corporate responsibility," which made corporations accountable to multiple stakeholders.

Sensing public frustration with the narrow profit motive of U.S. corporations, the Business Roundtable, an association of close to 200 influential U.S. CEOs, recently abandoned its 1997 shareholder-primacy position and declared that "the paramount duty of management and boards of directors is to the corporation's stakeholders." "Society gives each of us a license to operate," declared Ginni Rometty, IBM's CEO. "It's a question of whether society trusts you or not."

But doubts have been expressed about whether corporations can be trusted to regulate themselves, even after their stakeholder-primacy declaration. In a recent book, *The Anarchy*, historian William Darlymple describes the history of the East India Company, the most successful and most ruthless start-up in history. "Yet if history shows anything," write Darlymple, "it is that in the intricate dance between the power of the state and that of the corporation, while the latter can be regulated, the corporation will use all the resources in its power to resists."

One of the formidable resources that corporations can marshal is that of corporate personhood, which gives corporations the same legal rights enjoyed by natural persons. In fact, under U.S. law, some essential rights of the 14th Amendment, which addresses equal protection of the laws, belong not only to U.S. citizens but also to corporations. This has far-reaching implications. For example, the U.S. Supreme Court ruled in 2010 that corporate funding of independent political broadcasts in candidate elections cannot be limited under the First Amendment because of corporate personhood. This had led to a significant flow of corporate funds into U.S. political campaigns—and money buys influence in politics.

But the 14th Amendment was passed in response to issues related to former slaves following the American Civil War. How it came to be interpreted to grant personhood to corporations is a long and convoluted tale. Many argue that corporations should not have the same rights as natural persons. As IBM CEO Rometty said, society offers corporations a license to operate, so it makes sense for society to define the terms of that license, including rights and responsibilities, the issue of corporate personhood, and the relationship between shareholders and other stakeholders. Perhaps the time has come to formally define the terms of the relationship between society and corporations via a constitutional amendment that explicitly addresses the rights and responsibilities of corporations.

Follow me on Facebook and Twitter. C

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

DOI:10.1145/3361696

You Can Publish It! (You Have To)

THE VIEWPOINT COLUMN "Online Voting: We Can Do It! (We Have To)" in the September 2019 issue is naïve and unscientific. Although the column is explicitly framed as a response to the scientific community of experts who explain the dangers of Internet voting, it does not actually cite any of the scientific literature Ms. Orman is claiming to refute.

The scientific community (the "9 out of 10 experts" she mentions) have published many articles and reports laying out the scientific basis for why online voting is inherently insecure (given any known or imminently foreseeable technology).[1–7] Yet Ms. Orman does not cite any of these scientific papers among the bibliographic citations in the References section of her column. Given that *Communication*'s Viewpoint format does not permit an extensive bibliography, she did not have room to cite all of references listed here,[1–7] but in a response to the scientific community it would have been appropriate to cite (and explicitly respond to the science in) at least some of them.

There are gaping technical holes at the core of Ms. Orman's proposal. She proposes to rely on Trusted Platform Modules (TPM) to secure the end-user devices; but TPM cannot possibly do that within any foreseeable future, for two reasons. First, TPM replaces your trust in the device with your trust in the holder of the signing key. Intel or Google or Samsung or Apple holds the signing key of your device; shall we let them choose who wins our elections? And even if we did—TPM has been around for 20 years and we still keep finding security holes in it; it's simply not trustworthy.

I won't even begin to explain why Blockchain doesn't solve online voting, since that is so well explained in the scientific literature.[1,2] So too is the immensely thorny problem of distributing digital credentials to all voters, which Ms. Orman ignores entirely.

Even if one regards her Viewpoint as a guide to the many difficult scientific challenges that must be overcome before it's safe to proceed with online voting, the concluding paragraphs are completely pie-in-the-sky. She presumes that we could have secure smartphones with trusted hardware and software if only the government would subsidize them; as if well-resourced, technically savvy corporations such as Apple and Google were not already busting their butts to make their phones secure and failing in any case. And Ms. Orman suggests, in the very last paragraph, that secure TPM+TCB+PKI+(new-standardized-markup-language) could all happen within five years, by 2024, and be widespread by 2028. That claim is where the essential unreality of this whole scheme becomes clear. With so many intractable scientific problems unresolved—as they are even by Ms. Orman's own analysis—it is irresponsible to suggest pilot projects in elections for public office within such a short timeframe.

References

1. The Myth of "Secure" Blockchain Voting. D. Jefferson, Oct. 2018; www.verifiedvoting.org/jefferson_themythof_secure_blockchainvoting/.
2. Securing the Vote: Protecting American Democracy. National Academies of Science, Engineering, and Medicine, Sept. 2018; https://doi.org/10.17226/25120.
3. Email and Internet Voting, The Overlooked Threat to Election Security. S. Greenhalgh, S. Goodman, P. Rosenzweig, and J. Epstein, Oct. 2018.
4. The Future of Voting: End-to-End Verifiable Internet Voting—Specification and Feasibility Study. Report of the U.S. Vote Foundation, 2015; https://www.usvotefoundation.org/sites/default/files/E2EVIV_full_report.pdf.
5. If I Can Shop and Bank Online, Why Can't I Vote Online? D. Jefferson, 2011; https://www.verifiedvoting.org/resources/internet-voting/vote-online/.
6. Recommendations Report to the Legislative Assembly of British Columbia. The Independent Panel on Internet Voting, 2014; http://bit.ly/2lHEDYS.
7. Security Analysis of the Estonian Internet Voting System. J.A., Halderman, H. Hursti, et al., 2014; http://bit.ly/2lUlzXf

Andrew W. Appel, Princeton, NJ, USA

Author's Response:

No research proves that online voting a priori defies security principles. The growing set of innovative tools and techniques for software verification, trustworthy identity credentials, and publicly verified computation argues the contrary. As in all practical solutions, there will be a trade-off between cost and security.

My perspective is that the balance point is rapidly shifting, and security researchers and professionals need to produce, critique, analyze, and verify high-assurance voting systems. The volatility surrounding these issues should not deter progress.

Hilarie Orman, Woodland Hills, UT, USA

Editor-in-Chief's response:

In an era of active election interference by foreign powers in the U.S. and many other countries, the importance of careful design, vetting, and validation of online voting systems can't be overstated. At the same time, U.S. voter participation in national elections (the presidential elections every four years) has been mired in the 50%–60% range for past 50 years, so the need for technology that could increase participation in democracy are also desirable! This is an important issue where the experts of the ACM have contributed greatly to understanding and public policy, and there is much more to be done.

Andrew A. Chien, Editor-in-Chief

ACM Must Maintain Profession Neutrality

Companies like Google are strong supporters of ACM, sponsoring ACM's A.M. Turing Award and encouraging its employees to become ACM members. But that support gives ACM a greater, not lesser, responsibility to maintain objectivity and neutrality. Consequently, I was dismayed to read Vinton Cerf's editorial "Polyglot!" (Sept. 2019), a thinly veiled laundry list of all the wonderful things Google can do: "Google speaks 106 languages … Google's language ability vastly exceeds my own … [Google] Assistant … Google Lens … Google Translate …" and even "Google Science Fair." Cerf lauds Google eight times, failing to mention any other organization even once.

Cerf, a luminary of our field, is free to serve Google as its "chief evangelist," as his byline notes. ACM should not allow itself to be used as its platform.

Jonathan Grier, Pikesville, MD, USA

Editor-in-Chief's response:

It's a good point that ACM aspires to balance coverage of advanced technologies from leading academic researchers, government researchers, companies, and other leaders around the world. This case was a failure of expediency and familiarity. Vinton Cerf's employer certainly has no monopoly on advanced technology in language translation (for example, Microsoft Translator, Amazon Translate, Baidu Translate) and image recognition (for example, SenseTime, Amazon Rekognition, Bing Visual search). We will continue to strive to do better!

Andrew A. Chien, Editor-in-Chief

Coming Next Month in **COMMUNICATIONS**

The Rise of Serverless Computing

Automated Program Repair

Rethinking Search Engines and Recommendation Systems

Q&A with Garth Gibson

OpenPiton: An Open Source Hardware Platform for Your Research

Hack for Hire

Plus the latest news about malevolent machine learning, regulating IT, and robots for space.

BLOG@CACM

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

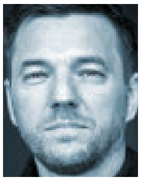
twitter

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

DOI:10.1145/3360907 http://cacm.acm.org/blogs/blog-cacm

The Benefits of Indolence

Yegor Bugayenko explains his realization that software developers should go neither above nor beyond.



Yegor Bugayenko
Lazy Developers Are the Best Developers
http://bit.ly/2lEC9KE
July 15, 2019

We are taught from a young age that the hardest workers enjoy the most success. Hard work pays off, or so we are told. But "hard work" can be a bit problematic for software developers, because it often means going well above and beyond the original scope of the project.

This is especially true when it comes to understanding legacy code. When you deal with legacy code, you often find yourself having to engage in so-called "deep thinking." You are expected to understand large problem scopes before you even begin trying to fix the small bugs. For a long time, this stressed me out. Then I got an idea: be lazy.

At my company, Zerocracy, we practice a #NoAltruism policy. We, quite literally, think only about ourselves and our personal profit. This might sound a bit harsh. Isn't it better to play nice and try to appease your clients? In an ideal world, maybe. But here's what we have learned about clients: they also practice #NoAltruism.

Clients want to keep costs low, and if they can, they will pass costs onto outside companies. That's why we decided to "get lazy" and only do what we are paid to do. We won't go out of our way to improve a project, refactor, or fix code unless we are getting paid for it.

And when we find ourselves with a task in front of us and we don't understand how to solve it, we usually don't blame ourselves. This is especially true if the problem has something to do with legacy code. See, here's the thing: we weren't paid to understand the legacy code. We were paid to add a feature, solve a bug, or whatever.

Suddenly becoming experts in a project's legacy code would be outside the scope of our work, and since we're lazy, we're not going to venture outside of our assignment unless we're paid to do so. A project shouldn't expect you to be intelligent or tech-savvy, as far as the legacy code is concerned. Instead, you need to focus on closing tickets.

It's not your fault if the code is a complete mess, or the bug is serious, or you can't estimate how much time it will take to understand the legacy code, let alone how to fix the bug. So whose fault is it? The first guilty party is the code itself. And the clients overseeing the code are also at fault.

Once you accept that, you can put together a basic report by creating new tickets. This report could be lazy-simple:

- There is no documentation for Class Y, can't figure out how it works.
- Library Z is in use but why aren't you using library B?
- This algorithm is a complex mess, can you explain what it does?
- The class naming rules are incoherent, can you provide documentation?

Suddenly, your initial "report" is instead a list of questions. You can't provide the answers because you don't honestly know them and you are too lazy to figure it out. Answering these questions falls outside of the scope of work you were hired for, so it is reasonable to expect the client to provide documentation.

Now, you might have noticed a common thread in the questions here. I didn't ask for help. I didn't ask someone to create something for me. Programmers will often reach out for help, saying something like "which library should I use for this task?"

Here's the thing: your clients aren't hiring you so they can do your work for you. They aren't hiring you so they can be your teacher, either. They don't really want to explain anything to you. For them, it's money and time they would rather not spend.

So your goal, then, is to get your clients to fix the code base so that the code itself becomes more obvious and easier to read. This will help not only you, but everyone else. As such, focus on asking for documentation and code source fixes.

Okay, so you've got the tickets out and you've asked the client to fix their source code and address other problems. So what now? Sit back and relax! You wait for the tickets to be resolved and don't sweat who is resolving the issues; that's not really our business.

Now, your employer may decide to kick the problem back to you, asking you to solve it on your own. That's fine, so long as you're getting paid for it and the employer expands the scope of your work. Instead of fixing bugs, you're now documenting some functionality or refactoring this and that.

As you create tickets and blame everyone else around you, you will continue to create smaller and smaller scopes. Eventually, you may find that the tickets can be fixed in a half hour or less. And keep in mind, when I say "blaming everyone else," that doesn't mean shouting at other people. It simply means not beating yourself up for problems you didn't create, and shifting responsibility for poorly written code to the original source.

Being lazy can take a lot of effort (seriously). We are programmed not to be lazy. Some people will resist the call. They might feel ashamed (stop it!). They want to be perfectionists (only perfect what you're paid to!). Or maybe you lack the passion needed to be lazy (get a new job!).

> **"It's not your fault if the code is a complete mess, or the bug is serious, or you can't estimate how much time it will take to understand the legacy code, let alone how to fix the bug."**

Comments

This is unacceptable practice from ACM's professional ethics guidelines. Zerocracy promotes no altruism and no help. This practice violates the core mission of ACM as an organization, which is "Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing." I request ACM to retract this article. Computing professionals have the obligation to behave in an altruistic manner and help each other for both advancement of business productivity, human well-being, and advancement of computing systems. Zerocracy is a disgraceful movement for computing profession.

—Mehmet Suzen

Mehmet, can you please elaborate on how exactly "contribute to society" leads to the conclusion that we are obliged to behave in an altruistic manner?

—Yegor Bugayenko

I think this policy is created to end the abuse on the client's behalf. #NoAltruism does not mean that in Zerocracy people would create software to support terrorism. Engineering is not altruistic, is precise. The Zerocracy policies are meant to create an efficient culture, not people without values. I think Mehmet misunderstood what #NoAltruism means.

—Eduardo Portal Geroy

"Computing professionals have the obligation to behave in altruistic manner and help each other for both advancement of business productivity, human well-being, and advancement of computing systems." As much as they have an obligation to not waste their time for free, increasing the engineering level in the company, helping others do their job, and saving time to help others and contribute to society in their free time, doing really altruistic things, not what you are talking about.

—Nikita Puzankov

"Computing professionals have the obligation to behave in an altruistic manner and help each other."

There's a difference between purposeful altruism as a means to improve the system, and blind altruism as a fanatic ideology. The thing we need to keep in mind is, the human psychology is never without its flaws, no matter how hardcore a saint you would be trying to play here. I myself have seen numerous examples of a biased altruist doing much more harm than a selfish but rational person in a similar situation.

Zerocracy is about regulating those psychological flaws, not trying to abolish them, which would most certainly end in (yet another) wasted effort. Being truthful with oneself, first and foremost, is the key in building all sorts of constructive professional relationships. Ignorance of that is bound to amplify guilt and fear in performers many times in the end, which might be appealing to certain moral fundamentalists who believe a scared programmer's guilt complex is like some sort of a virtue. The truth is, it just doesn't work out like that.

—Ilyas Gasanov

This occurs whether one is a consultant or contractor, or a salaried employee of the organization that owns the software. I have been both.

Even within the organization that owns the software, the deep thinking required to document otherwise undocumented systems or to fix underlying design problems is discouraged, and the attitude of "fix the immediate problem" prevails. This causes the organization's maintenance costs to increase steadily over time as technical debt piles up unaddressed, deeper and deeper.

This works similarly to the principle of conservation of energy, which pops up in infinitely varied guises whenever one attempts to create a perpetual motion machine: it is always thus regardless of which trendy or modern "methodology" is used in an attempt to manage the problem solved without doing the actual, necessary work.

In the end, one is doing one's client or one's employer a disservice by not warning them that a failure to solve the deeper problems will cost far more in the long run than any immediate savings they will realize by ignoring those problems for the present.

—Robert Watkins

Yegor Bugayenko is founder and CEO of software engineering and management platform Zerocracy.

The Essentials of Modern Software Engineering

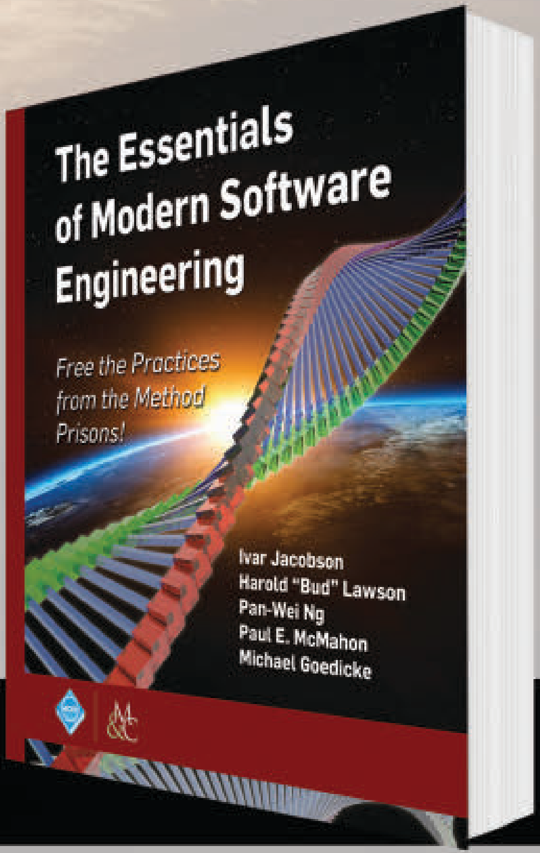
Free the Practices from the Method Prisons!

This text/reference is an in-depth introduction to the systematic, universal software engineering kernel known as “Essence.” This kernel was envisioned and originally created by Ivar Jacobson and his colleagues, developed by Software Engineering Method and Theory (SEMAT) and approved by The Object Management Group (OMG) as a standard in 2014. Essence is a practice-independent framework for thinking and reasoning about the practices we have and the practices we need. **It establishes a shared and standard understanding of what is at the heart of software development. Essence is agnostic to any particular methods, lifecycle independent, programming language independent, concise, scalable, extensible, and formally specified**. Essence frees the practices from their method prisons.

HIGH PRAISE FOR THE ESSENTIALS OF MODERN SOFTWARE ENGINEERING

“Essence is an important breakthrough in understanding the meaning of software engineering. It is a key contribution to the development of our discipline and I’m confident that this book will demonstrate the value of Essence to a wider audience. It too is an idea whose time has come.” – Ian Somerville, St. Andrews University, Scotland (author of *Software Engineering, 10th Edition*, Pearson)

“What you hold in your hands (or on your computer or tablet if you are so inclined) represents the deep thinking and broad experience of the authors, information you’ll find approachable, understandable, and, most importantly, actionable.” – Grady Booch, IBM Fellow, ACM Fellow, IEEE Fellow, BCS Ada Lovelace Award, and IEEE Computer Pioneer


The Essentials of Modern Software Engineering
Free the Practices from the Method Prisons!
Ivar Jacobson
Harold "Bud" Lawson
Pan-Wei Ng
Paul E. McMahon
Michael Goedicke

Ivar Jacobson, Harold “Bud” Lawson, Pan-Wei Ng, Paul E. McMahon, Michael Goedicke
ISBN: 978-1-947487-24-6
DOI: 10.1145/3277669
http://books.acm.org
http://store.morganclaypool.com/acm

ACM BOOKS
Collection I

Science | DOI:10.1145/3360909

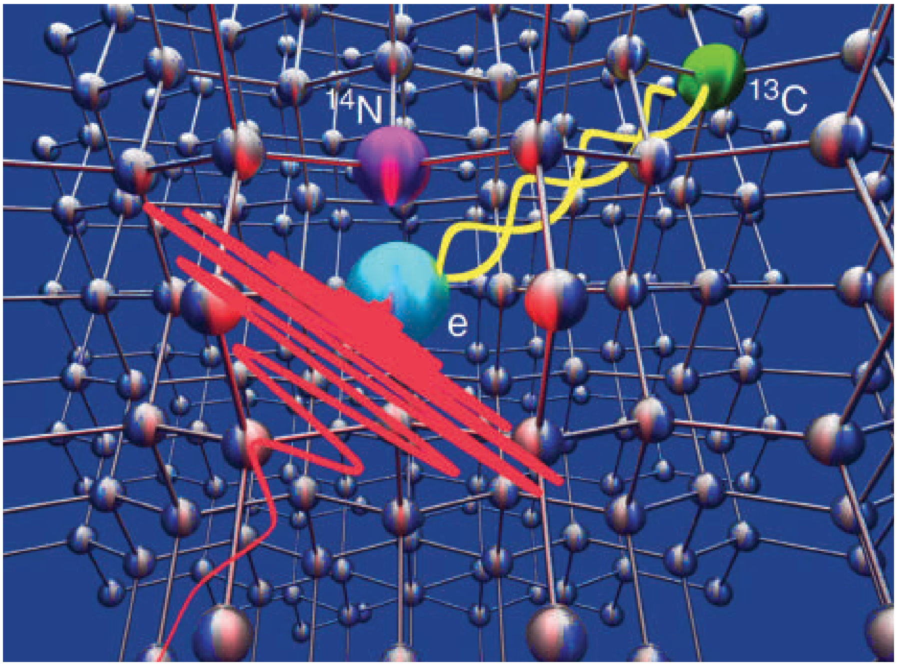
Don Monroe

Information Is Physics

Individual bits of information can have direct physical consequences.

EFFICIENT ERROR-CORRECTING CODES for quantum computing recently emerged from mathematical models used to study black holes. This surprising finding joins to a long list of profound connections between information and physics.

The most intriguing examples began as paradoxes or "thought experiments" that are hard to test experimentally. Physicists take them seriously because they challenge core concepts and may require revolutionary theoretical changes that could have practical consequences.



The lattice structure of carbon atoms in a diamond crystal contains a nitrogen-vacancy center with surrounding carbon nuclear spins. Researchers have demonstrated reliable quantum state transfer of photon polarization into a carbon isotope nuclear spin coupled to the nitrogen-vacancy center, based on photon-electron Bell state measurement by photon absorption.

IMAGE BY HIDEO KOSAKA; COMMUNICATIONS PHYSICS, 2, 74 (2019). DOI: 10.1038/S42005-019-0158-0

The Physics of Computation

The first hints that information has physical significance emerged in the 1800s, as researchers connected the somewhat mysterious thermodynamic quantity known as entropy to the information needed to describe a particular physical configuration. In this view, the progressive loss of information about an orderly initial state leads to the inexorable increases in entropy of an isolated system demanded by the Second Law of Thermodynamics, which constrains the efficiency of engines.

Beyond such statistical accounting, individual bits of information can have direct physical consequences, as illustrated by thermodynamics pioneer James Clerk Maxwell. He suggested that a "demon" that could see approaching molecules could merely open or close a trapdoor between two compartments of gas to let slow and fast molecules accumulate on opposites sides. The resulting temperature difference, seemingly without any entropy increase elsewhere, would violate the Second Law. (Maxwell's tricky creature also inspired the name of programs that operate behind the scenes in some operating systems.)

Physicists resolved the paradox by noting that Maxwell's demon eventually would need to erase the information it had gleaned about the molecules, and that this erasure would create enough entropy to preserve the Second Law. Overcoming the erasure of one bit requires an energy expenditure of a few billionths of a picojoule (at room temperature). This puts a lower bound on the energy needed for computation, because the output of a logic gate usually compresses the information in its inputs. Fortunately, current electronics devices use millions of times more energy per operation, so the limit is not (yet) important practically.

Quantum Information

Even this fundamental limit could in principle be avoided, however, by making all computations reversible, retaining enough information to reconstruct the original input. For researchers working on candidate components for quantum computers, this turns out to be immediately relevant, because these devices always operate reversibly, and this needs to be incorporated in circuit design.

Indeed, according to quantum mechanics, the mathematical evolution of any system is restricted to "unitary" transformations, which "basically means that whatever information you have, it always is there in some form," said theoretical physicist Brian Swingle of the University of Maryland. "Maybe it's very hard to read it out in some sense, but it's there."

Quantum information should still be conserved even when it is scrambled by interactions with the environment, which can be viewed as a larger quantum system. Although such interactions are often viewed as uncontrolled noise that causes "decoherence" that scrambles quantum information, quantum error-correction schemes exploit the overall reversibility of the combined system to ensure the desired information is preserved where it is needed.

Actually, "classical" (non-quantum) physics also follows microscopic equations that pay no attention to the direction of time. Indeed, physicists have long struggled to describe how such deterministic processes lead to the apparent loss of information embodied in increasing entropy, since the final state contains all of the details needed to reconstruct the initial state. "Information is just as preserved classically as in quantum," said Sean Carroll, a theoretical physicist at the California Institute of Technology (Caltech) in Pasadena, CA.

Beyond the Horizon

It is in the quantum realm, however, that information has raised the most profound conceptual challenges. This is most apparent in the field of quantum gravity, which aims to reconcile quantum mechanics and general relativity.

Traditionally, quantum mechanics plays out on a "stage" of unchanging spacetime, Swingle said. "If you try to make that stage dynamical, as happens in general relativity, where the geometry of spacetime is changing as a function of time, then combining those two things is hard."

The gravitational and quantum frameworks can usually agree to disagree, since they apply to very large and small scales respectively. However, their conflict becomes unavoidable for physicists studying black holes, which are both extremely massive and relatively compact. In this reconciliation effort, information plays a central role.

Once anything falls within a black hole's "event horizon," from which even light cannot escape, it should have no more influence on outside space. In particular, any information embedded in the infalling material is forever inaccessible. However, in the 1970s, Stephen Hawking of the U.K.'s University of Cambridge suggested that normally ephemeral pairs of particles that appear in the quantum-mechanical vacuum could be ripped apart at the horizon, with one sucked inside and the other escaping.

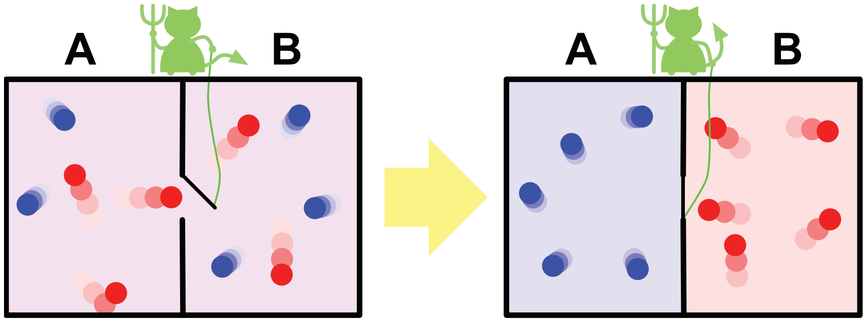
One consequence of this escaping "Hawking radiation" is that the black hole will eventually evaporate completely. At that time (usually ridiculously far in the future), information that had been carried in would not just be inaccessible, but gone forever, violating the quantum rule that it is always preserved. Physicists argued about how to resolve this "black-hole information paradox" for decades, but they largely came to accept that the information was somehow carried away in quantum "entanglement" between different radiated particles. In 2004, Hawking famously agreed, conceding a bet with the California Institute of Technology's John Preskill.

The Universe as a Hologram

Important support for this consensus came from a tool proposed in 1997 by Juan Maldacena of the Institute for Advanced Study in Princeton, NJ. It is called the AdS/CFT correspondence because it allows a mathematical mapping between a particular model of spacetime (AdS) and a class of quantum models (CFT).

Intriguingly, although the gravitational and quantum systems are equivalent, the quantum system has one fewer spatial dimension, somewhat like the surface of the gravitational system. This is an example of a "holographic" universe, so called because it resembles the way that a flat holographic film can encapsulate a three-dimensional image. Black holes constructed in this idealized universe can evaporate while conserving information. Carroll said the hope is that looking at such explicit examples will "reveal general principles," although he stresses that other approaches should also be explored.

The AdS/CFT framework has also yielded other insights, including new ways to study complex quantum-mechanical systems like superconductors by looking at the corresponding gravitational model. It also re-



A schematic of the Maxwell's demon thought experiment.

SCHEMATIC BY HTKYM/WIKIPEDIA (CC BY-SA 3.0)

vealed the surprising quantum error-correcting codes referred to at the start of the article, in which the surface information captures a subtle redundancy in the way information is encoded in the higher-dimensional bulk.

A Cosmic Firewall?

Hawking's concession did not end the controversies about black holes, because of ambiguity when one particle of an entangled pair falls in. Traditionally, this particle would remain entangled with its partner outside, sharing its quantum information. If this information is also carried away in the Hawking radiation, however, that violates a quantum rule known as "monogamy of entanglement."

For many years, physicists thought this conflict might be tolerable because no one could ever compare the information inside and outside, but in 2012 a group of physicists showed that this loophole fails right at the horizon. Instead, they proposed that passing through the horizon destroys the entanglement, creating a huge sheet of energy at the horizon known as the "firewall." This idea is repellent to many physicists, because a guiding principle for Einstein was that falling freely through space feels the same everywhere, with "no drama" at the event horizon. Other researchers have proposed other ideas, for example that the information carried in is conveyed to a different part of space by a "wormhole" and thus survives the evaporation.

These radical ideas, driven by the information paradox, threaten to restructure fundamental aspects of how physicists understand the universe. Although there remains no consensus on the resolution, Carroll said, it has "settled into something that many people agree is a problem."

Emergent Spacetime

As if wormholes were not exotic enough, Carroll, Swingle, and other physicists are exploring the idea that the entire structure of spacetime emerges from entangled quantum information. This alternative approach, sometimes called "It from Qubit," starts with abstract points, with no sense of space between them at all, said Swingle. "Then you start entangling them in some characteristic pattern, and that pattern can take on a geometric structure, in that you follow a link from one particle to another particle, eventually you have some sense of being able to go somewhere, some sense of distance, some sense of space."

Carroll, Swingle, and other physicists are exploring the idea that the entire structure of spacetime emerges from entangled quantum information.

This ambitious scheme remains a work in progress and may not prove successful, but there is no doubt that information will continue to guide fundamental thinking about physics. "Taking an information-theoretic point of view," Swingle said, can provide a "unifying framework to think about lots of different things. It's sort of a software versus a hardware view of the world."

Further Reading

Moskowitz, C.
Tangled Up in Spacetime, *Scientific American*, Oct. 26, 2016
http://bit.ly/2K5Oj87

Wood, C.
Black Hole Firewalls Could Be Too Tepid to Burn, *Quanta Magazine*, Aug. 22, 2018
http://bit.ly/2SzYaH3

It from Qubit: Simons Collaboration on Quantum Fields, Gravity and Information, Simons Foundation, http://bit.ly/32QylY7

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

ACM Member News

FROM VIDEO GAMES TO RISC-V ISA

"I got into computers through video games, like Space Invaders," says Krste Asanović, a professor in the computer science division of the department of electrical engineering and somputer sciences at the University of California, Berkeley (UC Berkeley). "I taught myself to program mainly to write games."

Asanović earned his bachelor's degree in electrical and information sciences from the University of Cambridge in the U.K., and received his Ph.D. in computer science from UC Berkeley.

While his main areas of focus are computer architecture, VLSI design, parallel programming, and operating system design and security, Asanović's recent focus has been on the RISC-V Foundation. RISC-V is a free and open Instruction Set Architecture (ISA) serving as the interface between hardware and software.

"The RISC-V ISA started at Berkeley," explains Asanović. "It is meant to replace Intel and ARM ISAs. It has lots of worldwide interest now."

The dominant industry ISAs are proprietary, Asanović says. Servers, desktop, and laptop computers are mostly built around Intel's x86 ISA, while mobile devices are built on ARM ISAs.

The point of RISC-V, Asanović says, "is to let anyone build their own processor and take advantage of its inherent design flexibility."

In terms of where computer architecture research is headed, Asanović thinks the exciting areas to explore are artificial intelligence (AI) and security. He feels new AI applications are creating incredible demand, while in security the situation is bad and getting worse.

"One of the challenges will be to create a truly secure computing environment," Asanović says.

—John Delaney

Technology | DOI:10.1145/3360913

Samuel Greengard

When Drones Fly

Drone technology is poised to enter the mainstream of business and society, but engineering robust controls remains a challenge.

As drones have matured into smarter and more practical machines, they have hummed, buzzed, and whirred their way into industries as diverse as movie production, agriculture, civil engineering, and insurance. It is entirely clear that autonomous drones will play a prominent role in business in the coming years. Firms such as Amazon, FedEx, and Uber have experimented with the technology to deliver packages, food, and more, while military agencies, emergency responders, gaming companies, entertainment firms, and others have explored other possibilities.

"Drones introduce far more efficient ways to accomplish some tasks," says Todd Curtis, president of Airsafe.com, a site that tracks drone and other aeronautic technologies.

Powering more advanced drones are more sophisticated on-board sensors and processors, better artificial intelligence (AI) algorithms, and more advanced controllers and communication systems. In addition, engineers are packing greater numbers of sensors into drones—and using them in different combinations—to create greater "awareness" of the surrounding environment. This sensing, when combined with GPS and other navigation capabilities, allows drones to tackle more advanced autonomous tasks, including devices that explore caverns or other hard-to-reach spaces, as well as underwater drones that conduct research by scanning oceans.

Yet, despite rapidly evolving capabilities, it also is clear that autonomous drones have not completely mastered the art and science of navigating and accomplishing their designated task. Buildings, birds, power lines, trees and people remain formidable obstacles for autonomous Unmanned Aerial Vehicles (UAVs), as they are known. Fog, snow, smoke, and dust present additional challenges.

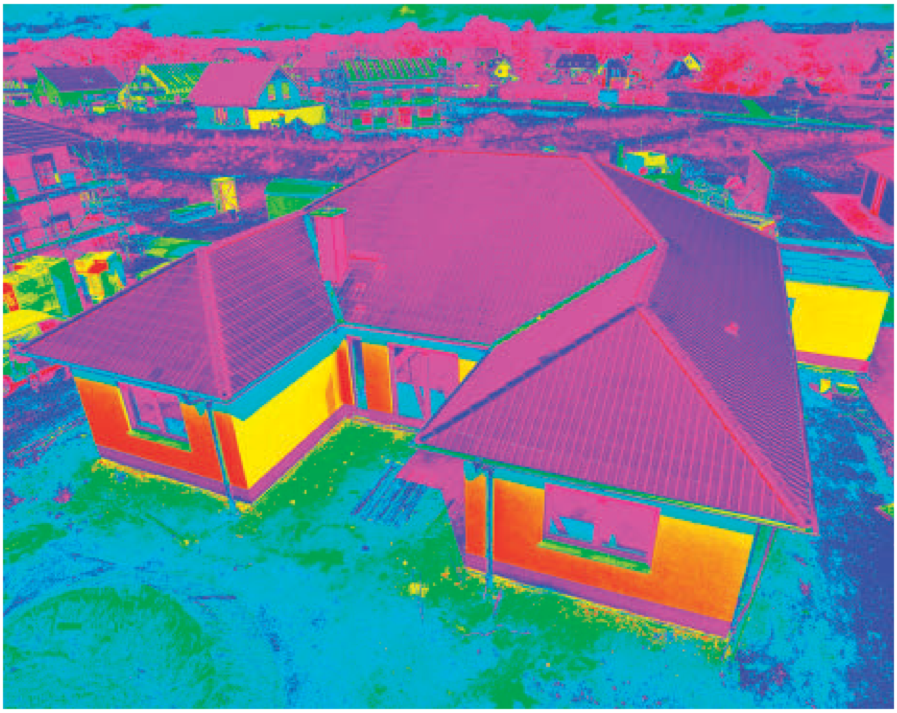


IMAGE BY RIKOBEST

It is one thing to showcase a drone in a controlled environment; it is quite another to have it operate flawlessly in the wild. UAVs must have near-perfect vision and sensing, as well as the ability to navigate areas where satellite and communications signals cannot reach and need backup and fail-safe systems that can take control of the drone if/when something goes astray.

"We are seeing remarkable advances in onboard sensing and processing, but also the use of far more sophisticated AI (artificial intelligence) algorithms in drones," says Nathan Michael, associate research professor at the Robotics Institute of Carnegie Mellon University. "These navigation and control systems are moving drones beyond the basic ability to fly from Point A to Point B. They're making it possible for drones to understand the world around them and make complex decisions in real time."

Drones Take Flight

Engineering a fully autonomous drone is rife with challenges—particularly in busy and complex urban areas.

First, they are not like the autonomous vehicles that operate on land. UAVs have extreme space and weight restrictions. Whereas a car can potentially have dozens, even hundreds, of sensors mounted across its surface, a drone can accommodate the weight of only a few.

Second, UAVs move in almost every direction in a three-dimensional (3D) space, while a motor vehicle operates on a two-dimensional plane. This makes designing software and algorithms for UAVs exponentially more complex.

Finally, the simple fact these machines are suspended in the air and constantly moving introduces additional challenges and risks.

Today, most UAVs operate on a line-of-sight basis. Essentially, a person uses a transmitter, typically operating in the 2.4GHz frequency band, to communicate with and control the drone's onboard computer. However, for drones to become truly autonomous, operate at high speeds, and ultimately become a commercially viable tool, onboard systems need to operate independently of humans (at least the vast majority of the time). This requires a dozen or more onboard sensors, such as cameras that work in both the visible and infrared spectra, LIDAR (light detection and ranging), or multi-spectral cameras; more advanced algorithms for understanding a wide range of environmental conditions; and sophisticated navigation systems that allow UAVs to sense their position more precisely.

There is also a need for improved safety systems—particularly in crowded urban areas. "Currently, drone companies add redundant propellers to avoid crashing. More advanced technology is necessary," says Davide Scaramuzza, director of the Robotics and Perception Group at the University of Zurich in Switzerland.

At drone manufacturing firms and in research labs, the next generation of drone controls and navigation systems is taking shape. Engineers and computer scientists are taking aim at various challenges, including how to process visual information at speeds reaching near 100 mph (160 kph), how to teach UAVs to react to unknown obstacles, what to do if the drone does not know how to respond to a given situation, and how to take over the controls for malfunctioning, rogue, or dangerous drones that may pose a threat. Not surprisingly, many of the decisions involve trade-offs. For example, it is already possible to fly an autonomous drone that has a very low probability of colliding with objects or crashing—as long as it flies at a very slow speed.

At the center of the challenge is simultaneous location and mapping (SLAM). Eric Amoroso, cofounder of KEF Robotics, a drone company that captured first place in a qualification round for a 2019 Lockheed-Martin UAV challenge, says inaccuracies in sensing and processing algorithms necessitates multiple onboard systems—as many as a dozen conventional cameras, vision sensors using such technologies as SWIR (short-wave infrared), MWIR (medium-wave infrared), LWIR (long-wave infrared), LIDAR (light detection and ranging), and radar (radio detection and ranging)—to robustly "see" what is going on around the drone.

What comes naturally to pilots when watching a UAV video stream is considerably more difficult for today's smartest UAVs.

What comes naturally to pilots when watching a UAV video stream—depth of field and localization of both static and dynamic objects—is considerably more difficult for today's smartest UAVs. Consequently, researchers are continuing to experiment with different combinations of sensors and SLAM algorithms to guarantee sight in cluttered environments. This includes stereoscopic vision and associated algorithms that help a drone gain depth-of-field and better understand relationships between and among objects—including other moving drones.

Equipping drones with vision and sensing capabilities that operate at the speed of flight is only part of the navigation challenge, however. There is also a need to ensure that a drone can process visual images quickly enough and make intelligent decisions in real time. Microprocessor and component manufacturers have introduced highly specialized chips that use increasingly powerful graphics processing units (GPUs) and accelerator chips to reduce visual processing time to milliseconds. Yet, further improvements are needed. For now, pilots can detect operational anomalies and react more quickly than an autonomous UAV. The ultimate objective for drone manufacturers is to push the devices' reaction time to the level of professional pilots so they can perform on par with humans, or perhaps even exceed them.

Machine learning will certainly make UAVs smarter and more agile, but it cannot completely solve the speed and latency problem. Moreover, better algorithms cannot anticipate every possible scenario or obstacle the world can toss at a drone. Ultimately, a UAV must be able to react to external events and avoid collisions while staying on course and accomplishing its intended task. Says Amoroso, "While a drone will likely not have the understanding a pilot has of the behavior of everyday objects, it nonetheless must react appropriately and quickly to avoid situations where it can cause harm to others or itself. Maybe the drone doesn't understand that branches can fall, or doors can open, but if given a robust enough SLAM system, it will still be able to navigate itself safely under such environmental disturbances."

Gaining Direction

Although GPS technology allows most drones to operate effectively most of the time, a dependence on satellites is not ideal—or even adequate—for companies looking to use UAVs for specialized commercial purposes. Objects such as buildings, trees or mountains might temporarily block signals. GPS also doesn't deliver the level of performance and precision needed when many drones operate autonomously close together. Without additional vision sensors and on-board navigation systems, collisions could occur, or drones might simply cease doing what they are supposed to do.

More advanced UAVs now incorporate a technology called Visual Inertial Navigation System (VINS) assistance. These systems rely on onboard cameras and inertial measurement units (IMUs) to track a drone's location when GPS signals are weak or nonexistent, such as in caves or deep valleys. Essentially, they work by detecting and tracking interest points across images and using them as anchor points for the robot to orientate itself, Scaramuzza says. In a certain sense, it's the drones mapping territory and using the map as they move over land, within caves, or underwater. However, this, too, has limitations since some environments change quickly.

Completely autonomous drones would require a combination of sensors, navigational capabilities, and communications links that push beyond current technology. They may also require new battery recharging systems—on the ground and in flight. Experts believe truly independent UAVs will take to the skies within the next few years, as further advances in computing hardware and software take place. Yet, in some cases, keeping humans in the flight loop may be desirable. This would likely include dangerous situations such as transporting a bomb, sending a drone into an unknown space such as a subterranean environment, or managing swarms of drones in highly cluttered airspaces.

Then there's the need to create failsafe systems to prevent UAV crashes. One solution, Amoroso says, is installing anomaly detection systems that alert a human to intervene when the drone can't navigate or operate normally. Another approach would be to place emergency beacons in commercial drones; if the UAV bumps into an object, it generates an alert or notification. Still another remedy, Curtis says, is programming malfunctioning drones to head to a safe space or simply to land until they can receive further instructions. Regardless of the specific approach, Carnegie Mellon's Michael says that any procedure leading to a human taking control of the system must be very well thought out. "Relying on a human to suddenly make an instantaneous decision could lead to potentially unsafe results," he cautions.

Yet the field is advancing, and even taking new directions. At the University of California Riverside, researchers have experimented with combined cellular signals and Wi-Fi to augment or replace satellite signals. At the Massachusetts Institute of Technology's Computer Science and Artificial Intelligence Laboratory, researchers are using virtual reality to train drones, and are build more robust algorithms by running virtual drones through simulations. Another team at the university has produced a mapping system called NanoMap that uses a depth-sensing system to stitch together ongoing measurements of the drone's immediate surroundings. This allows a single UAV—and theoretically a team of drones—to not only adapt motion and movement within a current field of view, but also anticipate how tormove in the hidden fields of view that it has already encountered.

DARPA is working on UAVS that will use sophisticated onboard mapping technology to remember places and things they have encountered.

Meanwhile, the U.S. Defense Advanced Research Projects Agency (DARPA) is working on UAVs that require no GPS, but fly at speeds up to 45 mph (72 kph). The devices will use sophisticated onboard mapping technology to remember places and things they have encountered. According to DARPA, the system could be used on the battlefield, and to rescue victims of natural disasters.

Into the Air

Researchers continue to explore ways to take autonomous drones to a higher level. This undoubtedly will revolve around better and more responsive cameras, faster and better image processing, and ongoing improvements in AI. For instance, Scaramuzza is focused on developing event-driven cameras with bio-inspired vision sensors that see only the motion in a scene. These smart pixels would reduce the processing load on the drone and allow it to focus on only the most important motion and activity. It would deliver high dynamic range at low power, even in low light conditions, while greatly reducing motion blur and latency. "I foresee that drones will become smarter and smarter and more and more situationally aware," he says.

Blending and optimizing existing technologies—and using increased processing power, better batteries, and improved algorithms, will result in additional gains, Michael argues. Part of the solution might also include mesh communication networks that use the collective intelligence of the group to teach and update individuals. This might best be described as real-time and collaborative machine learning. "The more the drones fly, the more experience they acquire. The more experience they acquire, the more they become high-performance machines. This makes them better equipped to navigate and mitigate challenging conditions," he says.

"We're moving toward a level of sophistication where onboard sensing systems and machine learning will create an environment that make it possible to step beyond basic navigation and create machines that use deliberate and intelligent decision-making. These systems—including groups of drones—will improve and get smarter over time," Michael says. "We're approaching an inflection point where drones will move past the novelty stage and become another capable system that can be used for a wide variety of purposes."

Further Reading

Kamat, S.U., and Rasane, K.
A Survey on Autonomous Navigation Techniques, 2018 2nd International Conference on Advances in Electronics, Computers and Communications, IEEE.
https://ieeexplore.ieee.org/abstract/document/8479446

Simon, N., and Songmahadthai, D.
Multi-drone Control System, Mälardalen University School of Innovation Design and Engineering, Jan. 16, 2019.
http://www.diva-portal.org/smash/get/diva2:1292032/FULLTEXT01.pdf

Mozaffari, M., Saad, W., Bennis, M., and Debbah, M.
Communications and Control for Wireless Drone-Based Antenna Array, *IEEE Transactions on Communications*, Vol. 67, Issue 1, Sept. 20, 2018, pp. 820–834.
https://ieeexplore.ieee.org/abstract/document/8469055/citations#citations

Kim, J., Seokhwa, K., Jaehoon, J., Hyoungshick, K., Jung-Soo, P., and Taeho, K.
CBDN: Cloud-Based Drone Navigation for Efficient Battery Charging in Drone Networks, *IEEE Transactions on Intelligent Transportation Systems*, Dec. 12, 2018, pp. 1–18. https://ieeexplore.ieee.org/abstract/document/8574043/authors#authors

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

Society | DOI:10.1145/3360911

Logan Kugler

Real-World Applications for Drones

Unmanned vehicles have a number of compelling real-world use cases.

In June, Amazon announced it was close to being able to offer for package deliveries by drone for its Prime Air service. That same month, Uber said it plans to test food delivery by aerial drone in crowded cities. And drone delivery company Flytrex already touts the ability to deliver drinks via unmanned vehicle on the golf course.

Despite such announcements, drones are not crowding the skies over major cities and population centers just yet. But that may be about to change.

After several years of hype, widespread drone usage may be close to ready for primetime.

Drones increasingly are being deployed in a number of compelling real-world use cases. These use cases have drone companies and enthusiasts bullish that, no matter what happens, there are serious real-world applications for drone technology today and in the near future that will disrupt life and business as we know it.

IMAGE FROM SHUTTERSTOCK.COM

Drone-Assisted Photography/Surveying

"Traditionally, we've seen drones being used for photography and surveying," says Eric Peck, CEO of Swoop Aero, an Australian company that delivers medical supplies via aerial drone. "It's all about data capture, because data really is driving the ability to generate economic growth at the moment."

From construction to insurance to real estate to agriculture, the ability to survey and photograph wide swaths of land and hard-to-reach locations with aerial drones is valuable to companies. For instance, high-quality photos and videos from different aerial angles can better showcase residential properties up for sale, more effectively highlighting elements that appeal to buyers.

Aerial footage shot by drones is less expensive than manually taking aerial footage from a helicopter. One drone photographer interviewed by *The Baltimore Sun* noted the cost differences: "I can drive up to my destination, plug my equipment in, and be done [photographing] in five or 10 minutes," said Jack Hardway, owner of a drone photography firm. "It doesn't cost me $5,000. It costs me pennies to put that thing in the air."

The cost is one benefit. The ability to collect more visual data from more angles than from a traditional camera also is important.

A Santa Monica, CA-based company called DroneBase uses unmanned aerial vehicles (UAVs, or aerial drones) to offer, among other services, aerial surveying of building rooftops to give insurance companies an easy way to assess damage related to claims. For insurance and surveying purposes, aerial drones offer the ability to cover more ground while traversing more areas and angles than might be possible (or affordable) with traditional manned aircraft.

Other use cases include surveying and monitoring progress at construction sites, and performing simple regulatory inspections for commercial real estate properties. Aerial drones are even used to fly around warehouses and find supplies or products faster and more accurately than humans do.

Aerial drones also come in handy in agricultural applications. They offer a dual benefit in this context. First, drones are used to survey fields. Instead of having to traverse hundreds or thousands of acres on foot or by vehicle, farmers have the ability to fly drones faster and more efficiently over large areas. That helps reduce the time it takes to monitor fields, as well as reducing the amount of fertilizer and pesticides they must use to maintain crops.

"We identify diseases and pests and fungus and weeds in the crop at an earlier stage," U.K. farmer Colin Rayner told German broadcaster DW. Some drones are even used to spray fields with pesticides. According to DW, Chinese drone company DJI sold 20,000 pesticide-spraying drones in 2018 alone.

In all of these examples, the drones are being piloted remotely by experienced professionals. For instance, DroneBase claims it has the "largest network of professional drone pilots in the world," but they are all still human beings.

Right now, this gives an advantage to bigger companies that can scale and capture cost advantages that offset the expense of human pilots.

"While the market for drone photography and data capture is massive, it's close to saturated right now, both in terms of platforms and operators," says Peck of Swoop Aero. "We'll see a lot of movement as big players gain advanced regulatory approvals [for more extensive drone applications], which allow them to gain a cost advantage based on economies of scale and drive smaller operators out of the market."

That dynamic has led companies like Swoop Aero to look at use cases for drones that involve delivering high-value commodities.

Drone Doctors

One high-value commodity that makes a lot of economic sense to deliver via drone is medicine.

The market need is clear: when it comes to perishable medical samples or life-saving vaccines, time is of the essence, and few technologies are better at traversing crowded or hard-to-reach areas than aerial drones.

Swoop Aero operates drone networks that deliver medicine quickly in countries like Pacific island-nation Vanuatu, which is composed of dozens of islands. Often, the deliveries take a fraction of the time they would if conducted by boat. Vanuatu is a country in which, the United Nations Children's Fund (UNICEF) estimates, a full 20% of children under five do not receive all the vaccines they need because of the logistical challenges around medicine delivery.

Last December, Joy Nowai of Vanuatu was, according to the company, the "first child in the world to be vaccinated with a vaccine delivered by a drone under a commercial contract," thanks to Swoop Aero. The drone travelled 30-plus miles to deliver the vaccine, while keeping it at the optimum temperature during the entire trip.

After the successful delivery, the company conducted a further four-month trial in Vanuatu, which Swoop Aero says led to continued work with the country's Ministry of Health.

Swoop Aero is now preparing to deploy additional drone networks to countries that lack easy logistical ac-

ACM News

Digital Transformation: A Business Imperative

Digital transformation, in which companies utilize advanced technologies like artificial intelligence (AI), cloud computing, and the Internet of Things (IoT), has become a business imperative for virtually any organization that wants to stay competitive and meet customer needs.

Many organizations have moved beyond Digital 1.0, where improving the speed of response was the strategic imperative, to Digital 2.0 and being able to anticipate customer requirements, according to Jamie Snowdon, chief data officer at HfS Research. "Having the right data to anticipate customer needs and support decisions that serve customers ahead of time delivers significant digital competitive advantages," Snowdon says.

Companies' chief information officers (CIOs) continue to drive digital transformation initiatives most often (28%), although CEOs are increasingly playing a leadership role in such transformation (23%), according to Altimeter's 2018–19 State of Digital Transformation.

Yet organizations often struggle with digital transformation initiatives, and have varying degrees of success. A 2018 Capgemini study of 1,300 executives found that only 39% have the digital capabilities required, and only 35% have the right leadership capabilities (mainly because digital isn't in their DNA).

In terms of the aftermath, a September 2018 survey by management consulting firm McKinsey of 1,733 executives involved in digital transformation efforts at their companies found only 3% of respondents have had complete success at sustaining their digital efforts.

Some 89% of information technology (IT) decision makers said their digital innovation investments have been "moderately or very successful," according to Insight's 2019 Intelligent Technology Index.

Nevertheless, organizations are forging ahead, and those with well-defined objectives and goals and a digital mindset are reaping the benefits.

JetBlue Airways, for example, is expanding its brand beyond travel and reinventing itself as "a tech company in the customer service business,'' said Eash Sundaram, JetBlue executive vice president and chief digital and technology officer, during remarks at the Massachusetts Institute of Technology CIO Symposium in May. The airline's customer strategy has become "personal, helpful, simple," and "not more tech; better tech," Sundaram said.

That sentiment was echoed by Kris Rao, CIO of Ricoh USA, who said his firm is moving away from its roots as an office equipment provider to become a digital company "empowering digital products." Reaching across the aisle [from within the silo of IT to the business] is the job of a technologist," said Rao, who also spoke at the Symposium.

McKinsey's April 2019 survey found organizations reporting the greatest levels of success in their digital transformations "ruthlessly focus on a handful of digital themes tied to performance outcomes." Additionally, these organizations "boldly establish enterprisewide efforts and build new businesses."

In addition, McKinsey found such companies create an adaptive design that allows for flexibility in the transformation strategy and resource allocation, and adopt agile execution practices and mindsets by encouraging risk taking and collaboration across parts of the organization.

Perhaps most importantly, the management consulting firm wrote in its summary of results, "In successful efforts, leadership and accountability are crystal clear for each portion of the transformation."

—Esther Shein is a freelance technology and business writer based in the Boston area.

cess to life-saving medicines and vaccines. One of the company's networks is being deployed in the Democratic Republic of the Congo, in collaboration with that country's Ministry of Health and local non-governmental organizations.

Another company, California-based Zipline, just launched a series of drone distribution centers in Ghana designed to deliver vaccines and medications to the country's population 24 hours a day. The company says health workers placing orders for medications via text can expect the requested medicine to be delivered within 30 minutes.

This type of aerial delivery at scale is not only well suited for geographically inaccessible areas. Earlier this year, logistics company UPS, in partnership with California-based drone startup Matternet, launched a pilot program to deliver medical samples around the campus of WakeMed, a not-for-profit health care system in Raleigh, NC. According to *Business Insider*, such samples used to take up to 30 minutes to deliver due to traffic congestion, but using drones, the deliveries now take just over three minutes.

"Transport is a clear opportunity commercially," says Peck. "We are focused on last-mile logistics for high-value commodities, predominantly in healthcare. It's a market which is forecast to grow in size from close to zero right now, to be worth over $10 billion in the next seven years."

Drones are not just flying to areas where people find it hard or time-consuming to go; they are swimming there, too.

In 2018, submersible drones built by Texas-based Ocean Infinity worked together to survey parts of the ocean floor inaccessible to humans. The goal was to find the remains of ships that had gone missing. The company succeeded in discovering the wrecks of an Argentinian submarine and a South Korean commercial vessel, long after hope was lost that concerned parties would learn the fate of the disappeared craft.

Like their aerial counterparts, underwater drones are packed with sensors that collect data and share it with company control centers. Instead of putting humans at risk in dangerous underwater conditions, the drones go deeper underwater and stay under longer than human-manned vessels could, transmitting back valuable data all the while.

Companies are desperately trying to make it possible for consumers to receive deliveries by drone of products that would normally be delivered by mail or manned vehicle.

The unmanned underwater drone approach has been so effective that one company tasked with finding Malaysia Airlines Flight MH370, the Dutch geosciences company called Fugro NV, reportedly "plans to do away with some [human] crews entirely."

The Future of Drone Delivery

There are plenty of intriguing real-world applications for drones at present, but that has not stopped companies from salivating over the holy grail of drones: last-mile logistics. Companies such as Amazon are desperately trying to make it possible for consumers to receive deliveries by drone of products that would normally be delivered by mail or manned vehicle.

While some companies focus on delivering high-value goods via unmanned or semiautonomous drones, last-mile logistics at scale requires near-full autonomy. To deliver most or all products at scale, drones from a company like Amazon will need have the ability to fly themselves short distances with little to no human involvement.

That means building highly sophisticated, reliable types of artificial intelligence (AI) and machine learning into delivery drones. These AI-powered systems must be able to visually recognize and physically respond to real-world obstacles with a high degree of accuracy and speed.

That could take some time, Keith Lynn, a program manager for Lockheed Martin's autonomous drone racing competition, told *The New York Times*. "Right now, autonomous drones are a thing you'd only find in labs, being pioneered by a small, niche audience."

A big reason for that is because autonomous drones struggle to make sense of visual information, particularly at high speeds, in part because of shortcomings in the sensors they utilize. Also, the faster the drones fly, dive, or drive, the more difficult it is for today's algorithms and cameras to process images at the speed required to recognize (and avoid) obstacles.

Aero Swoop's Peck is optimistic advancements will lead to drones having greater autonomy in the near future. "Over the next five to 10 years, we are going to see increasing levels of full autonomy used across all aspects of aviation," he predicts.

Further Reading

Hamilton, I.
Amazon drone deliveries are coming, but Jeff Bezos still missed his own deadline for airborne logistics, *Business Insider*, Jun. 6, 2019, http://bit.ly/2kJWNcv

Holley, P.
Uber plans to start delivering fast food via drone this summer, *The Washington Post*, Jun. 13, 2019, https://wapo.st/2m46EK9

Thompson, F.
Next generation farming: How drones are changing the face of British agriculture, *DW*, Jul. 19, 2019, http://bit.ly/2kviDAk

Waseem, F.
Howard drone users search for opportunity as 'the skies open', *The Baltimore Sun*, July 7, 2016, http://bit.ly/2mit76q

Wilke, J.
A drone program taking flight, *Amazon*, Jun. 5, 2019, https://blog.aboutamazon.com/transportation/a-drone-program-taking-flight

Wise, J.
Underwater Drones Nearly Triple Data From the Ocean Floor, *Bloomberg Businessweek*, Jun. 7, 2019, https://bloom.bg/2kobrpq

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

ACM ON A MISSION TO SOLVE TOMORROW.

Dear Colleague,

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

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

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

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

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

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

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

Sincerely,

Cherri M. Pancake
President
Association for Computing Machinery

Association for
Computing Machinery

Advancing Computing as a Science & Profession

DOI:10.1145/3363179

Pamela Samuelson

Legally Speaking
Europe's Controversial Digital Copyright Directive Finalized

Considering the new liability risks for ISPs, search engines, and news aggregators under recent EU-wide mandatory rules.

INTERNET GOVERNANCE RULES in the EU are about to change radically. The final version of its Directive on Copyright and Related Rights in the Digital Single Market (DSM), which has been under consideration for the past three years, was promulgated on April 17, 2019. EU member states now have two years to transpose the Directive's rules into their national laws.

In some respects, the DSM Directive is better than previous drafts (of which more anon). There is still reason to worry the new rules will be harmful for freedom of expression and information privacy interests of individual creators and users. How much harm will depend on how member states implement the Directive and how courts interpret it, as many of its terms are ambiguous.

This column discusses key differences between earlier drafts of the DSM Directive and the final version and makes some general observations on some aspirations that underlie this Directive.

Repeal of the Safe Harbor for ISP Storage of User Contents

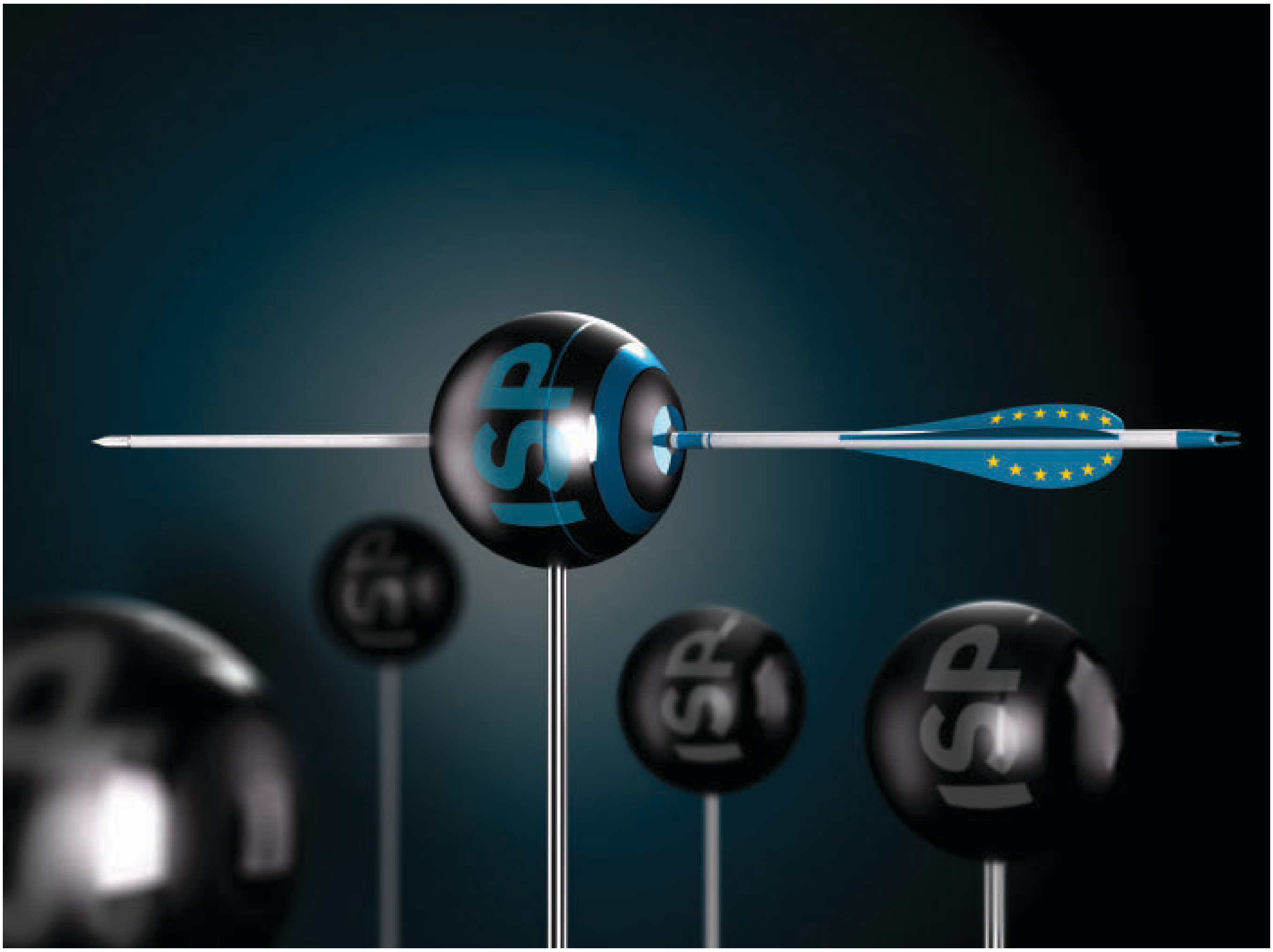
The most significant and controversial of the new DSM rules is the stiffer liability rules the Directive established for online content-sharing platforms, such as YouTube and Facebook.

There is still reason to worry the new rules will be harmful for freedom of expression and information privacy requests.

Under laws in place in the EU and U.S. since 1998, Internet service providers (ISPs) have enjoyed a safe harbor from liability for infringing acts of their users of which the ISPs were unaware. ISPs faced liability only if they failed to investigate and take down infringing materials after receiving notice from copyright owners about where such materials were located.

Article 17 of the DSM Directive (Article 13 under previous drafts) imposes strict liability on online content-sharing sites for user infringements and obliges them to use "best efforts to ensure the unavailability of specific works." Because EU member states may decide that "best efforts" requires platforms to use filtering technologies, this provision has often been called the "upload filter" rule. (Previous drafts of the Directive were more pointed about the need to use filtering technologies.)

There are two exceptions to the DSM



COMPOSITION BY ANDRIJ BORYS ASSOCIATES, USING IMAGE BY OLIVIER LE MOAL

Directive's strict liability rules for online content-sharing platforms. One is for nonprofit services, such as online encyclopedias, educational and scientific repositories, and open source software developing platforms. A second is for startup online content-sharing services that have been available to the EU public for less than three years and that have annual revenues of 10 million euros or less. The liability of these two types of services for user infringements are subject to compliance with the existing notice and takedown rules.

Critics have charged that the DSM strict liability rules will interfere with user freedoms to make lawful uses of copyrighted works, such as parodies or critical commentaries, because filtering technologies cannot distinguish between outright infringements and privileged uses.

Seemingly in response to this criticism, Article 17 now states that member states "shall ensure that users in each Member State are able to rely on exceptions for "(a) quotation, criticism, review; (b) use for the purpose of caricature, parody, or pastiche."

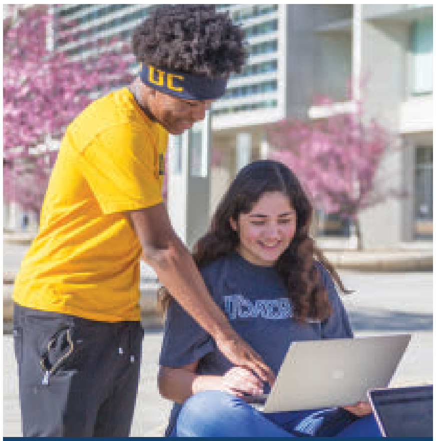
Whether this effort to ensure privileged uses can be uploaded to content-sharing sites will meaningfully limit the Directive's scope or serve only as aspirational window dressing remains to be seen. It seems unlikely, though, that EU member states can require developers of filtering technologies to refine their algorithms so that all parodies, critical comments, and other privileged uses will remain available to the public. Yet, this seems to be the only way to ensure privileged uses can be preserved.

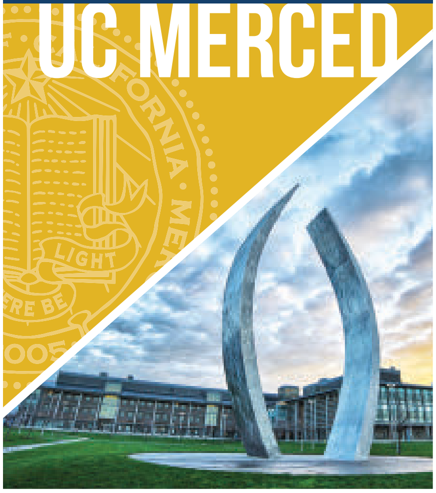
Text and Data Mining Exceptions

Under existing EU law, text and data mining on digital repositories of copyrighted works and databases had an uncertain status. The drafters of the DSM Directive decided this activity should be lawful because of the important insights the use of such research tools can enable. To this end, they proposed in Article 3 that member states adopt a mandatory new exception to copyright and database rules to allow nonprofit research and cultural heritage institutions to engage in text and data mining for scientific research purposes.

While this exception was good so far as it went, earlier versions of the DSM Directive would have left independent researchers and profit-making text and data miners out in the cold. Because EU policymakers aspire to foster the growth of artificial intelligence and other data-intensive businesses, they came to recognize restricting text and data mining to nonprofit scientific research was shortsighted, especially given that other countries, notably the U.S. and Japan, have adopted broader text and data mining privileges.

While Article 3 retains the text and data mining exemption for nonprofit scientific research, the final DSM Directive sets forth a new Article 4 requiring member states to create a more general mandatory exception to copy-





right and database rules to allow text and data mining by independent researchers and profit-making establishments without restriction on purpose.

Although Article 4 is broader than Article 3 in the users and uses to which it would apply, Article 4 is more limited than Article 3 in two respects: First, the Article 4 exception does not apply to the extent that rights holders have expressly reserved the right to control text and data mining. Second, the Article 4 exception can be overridden by contract, whereas the Article 3 exception is nonwaivable by contract.

Press Publishers Right

The final version of the DSM Directive directs member states in Article 15 (previously Article 11) to grant press publishers two years of exclusive rights to control reproductions and communications to the public by information society service providers.

Earlier versions of the DSM Directive's press publisher right attracted intense criticism. Opponents charged it would impede the free flow of news and other information vital to a democratic society, harm journalists who often rely on search engines and aggregators, and create uncertainty about its coverage and scope. Critics also thought this new right was unnecessary, unlikely to produce significant licensing revenues, and likely to further entrench powerful media conglomerates and global platforms to the detriment of smaller players.

Critics also expressed concern about how the new publisher right would interact with existing copyright laws, which typically allow for fair quotation rights, as well as with database rights, which allow users to extract insubstantial parts of databases.

Notwithstanding serious concerns about the press publisher right, the EU Council and Parliament decided to approve the grant of this new set of exclusive rights.

Earlier versions of the DSM Directive's press publisher right attracted intense criticism.

Seemingly to counter the charge that Article 15 would create a "link tax," Article 15(1) explicitly provides that the press publisher right does not apply to hyperlinking. In an effort to further narrow its reach, Article 15(1) says it would not apply to "private or noncommercial uses of press publications by individual users." Nor would it apply to use of individual words or very short extracts of a press publication."

But what exactly constitutes a "very short extract" of a press publication is unclear. Ambiguities about this and other terms in Article 15 makes it unlikely that member states of the EU will implement this new right in a harmonious way.

Licensing as a Goal of the DSM Directive

Proponents of the DSM Directive told European policymakers a powerful story in support of the new liability and exclusive rights rules that the Directive has now established. They assert there is a "value gap" the Directive could correct.

The short version of that story is that U.S. technology companies are making huge revenues from their uses of European rights holders' contents and too little of these revenues are flowing to European content providers. (Both European and American commentators have expressed considerable skepticism about the "value gap" story, but it was an influential part of the rationale for adopting Articles 15 and 17.)

To narrow, if not close, this gap, the DSM Directive aims to induce technology companies to negotiate for licenses. If such licensing occurs, then the stricter rules will not need to be applied, and worries about harms to freedom of expression and other social values expressed by critics of the stricter liability rules will not come to pass.

Consider, for instance, Article 17(1). After providing that online content-sharing sites will be strictly liable for giving the public access to infringing copyright-protected contents uploaded by users, that provision goes on to say that to avoid this liability, such sites "shall obtain an authorization" from rights holders "by concluding a licensing agreement" that would cover any otherwise infringing uploads.

It is not even remotely possible for online content-sharing services to get licenses from every copyright owner of European works available in digital form. The aspiration of Article 17 seems to be to induce platforms to obtain licenses from major European copyright sectors, such as motion-picture producers, recording-industry firms, and collecting societies that represent other kinds of rights holders (such as performing artists).

Article 17 gives European rights holders considerable leverage to insist on substantial revenue flows and other licensor-friendly terms as a condition of granting such licenses. Negotiating such licenses will be daunting because each member state of the EU has its own national law, domestic copyright industries, and collecting societies. Despite the Directive's aspiration to establish a "digital single market," no such market exists. YouTube and Facebook may be able to navigate the complexities of the EU markets, but smaller service providers may find it difficult or impossible to conclude negotiations that will shield them from Article 17 liability.

Licensing is also the principal goal of Article 15. The Recitals of the DSM Directive, which serve as a kind of explanatory preamble, emphasize that high-quality journalism, which is important to fostering well-informed public debate and democratic discourse, is expensive to produce. The goal of Article 15 is to enable licensing so that press publishers can develop sustainable business models.

Although news aggregators, monitoring services, and search engines make considerable revenues from advertising or subscriptions, very little, if any, of those revenues are shared with the press publishers, which seems unfair because the contents these services provide to their users come from those publishers.

As with Article 17, Article 15 creates a liability risk for online services that make use of EU press publisher contents that only licensing can overcome. As with Article 17, Article 15 provides press publishers with considerable leverage to conclude licenses on favorable terms to EU firms.

Whether the licensing goals of the DSM Directive will be fulfilled also remains to be seen.

Conclusion

It remains to be seen how EU member states will transpose the rules set forth in Articles 15 and 17 in their national laws. Perhaps some national legislators will coordinate efforts to resolve some key ambiguities in the Directive (such as the "best efforts" language of Article 17 and "small extracts" in Article 15) in a manner that will enable the relevant online service providers to assess the risks of liability and benefits of licensing on fair and reasonable terms.

Whether the licensing goals of the DSM Directive will be fulfilled also remains to be seen. Some online content-sharing sites may decide to license European contents, but many smaller entities may decide to risk liability and/or limit the availability of their services in the EU.

The experience of Germany and Spain, both of which adopted a press publisher right similar to Article 15, does not bode well. Both countries hoped to induce U.S. tech companies to license press publisher news these services provided to their users. Very few licenses were concluded, and some online services just stopped providing news from those countries.

Maybe the EU-wide nature of the new DSM rights will serve as a stronger incentive for licensing, but it is too early to conclude that either Article 15 or Article 17 will be effective in bringing more revenues to EU rights holders. One thing is for sure: U.S. online services providers face some difficult challenges in deciding how to proceed in response to the new DSM rules.

Pamela Samuelson (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley, and a member of the ACM Council.

Calendar of Events

October

Oct. 28–29
CSLAW '19: Symposium on Computer Science and Law,
New York, NY,
Sponsored: ACM/SIG,
Contact: Joan Feigenbaum,
Email: Joan.Feigenbaum@yale.edu

Oct. 28–30
ASSETS '19: The 21st Int'l ACM SIGACCESS Conference on Computers and Accessibility,
Sponsored: ACM/SIG,
Contact: Jeffrey Philip Bigham,
Email: jeffreybigham@gmail.com

November

Nov. 3–7
CIKM '19: The 28th ACM Int'l Conference on Information and Knowledge Management,
Beijing, China,
Co-Sponsored: ACM/SIG,
Contact: Wenwu Zhu,
Email: wwzhu@tsinghua.edu.cn

Nov. 3–6
SIGUCCS '19: ACM SIGUCCS Annual Conference
New Orleans, LA,
Sponsored: ACM/SIG,
Contact: Robert Haring-Smith,
Email: rharingsmith@alum.swarthmore.edu

Nov. 5–8
SIGSPATIAL '19: 27th ACM SIGSPATIAL Int'l Conference on Advances in Geographic Information Systems
Chicago, IL,
Sponsored: ACM/SIG,
Contact: Farnoush Banaei-Kashani,
Email: farnoush.banaei-kashani@ucdenver.edu

Nov. 9–13
CSCW '19: Computer Supported Cooperative Work and Social Computing
Austin, TX,
Sponsored: ACM/SIG,
Contact: Karrie Karahal,
Email: kkarahal@illinois.edu

Nov. 10–13
SenSys '19: The 17th ACM Conference on Embedded Networked Sensor Systems,
New York, NY,
Co-Sponsored: ACM/SIG,
Contact: Raghu Ganti,
Email: rganti@us.ibm.com

DOI:10.1145/3363181

Mark Guzdial, Alan Kay, Cathie Norris, and Elliot Soloway

Education

Computational Thinking Should Just be Good Thinking

Seeking to change computing teaching to improve computer science.

JEANNETTE WING'S 2006 *Communications* Viewpoint on computational thinking[5] ignited a worldwide movement to give students new knowledge and skills to solve problems in their daily lives. Quickly, teachers, curriculum and standards writers, and other education specialists were proposing what children needed to know about computation and how to develop a computational mindset. There is still little evidence that knowing about computation improves everyday problem-solving, but there is no doubt that Wing's call to action led to a broad and dramatic response.

The computational thinking movement puts the onus on the student and on the education system. They argue that if we change *humans* to think in ways that are informed by how we now work with *computers*, that will have problem-solving advantages for the humans.

Maybe.

If a city does not work for the residents, we could change the residents. Alternatively, we could redesign the city. The best urban redesign has citizens understanding the purpose and actively participating, so there is parallel development of both the city and the citizens.

Children today already think with computation. If we want better thinking and problem-solving, we have to improve the computing and use that to change our teaching. We put the onus back on the computer scientists and other computationalists. It is our job to design better.

IMAGE BY OLLYY

For Our Children, Computational Thinking Is Just Thinking

Tool use shapes thinking. While we might not think like a carpenter when we start using carpentry tools, if we apply ourselves (for example, reflect on our doing, as Dewey suggests[2]), we can develop carpentry thinking. We can learn to see what is possible with the tools of carpentry, the way a carpenter thinks.

Closer to home, the "kids these days" use all manner of digital—read: computational—tools. Before drawing the obvious conclusion, consider the following vignettes.

Vignette 1: Consider the following two problems, drawn from a research study[1]:

- (Algebraic Context): Given the following statement: "There are six times as many students as professors at this university." Write an equation that represents the above statement. Use *S* for the number of students and *P* for the number of professors.

- (Computer Programming Context): Given the following statement: "There are six times as many students as professors at this university." Write a computer program that will output the number of students when supplied (via user input) with the number of professors. Use *S* for the number of students and *P* for the number of professors.

While the equation in both problems is the same—*S=6P*—significantly more undergraduate engineering students provided the correct equation in the Computer Programing Context than in the Algebraic Context.

Vignette 2: Now, consider the following research finding (appearing in Norris and Soloway[4]):

- "[K–12] Students using word processors for writing generally produce longer, higher-quality writing than students using pencil or pen and paper."

The computational tool plays a role in students' ability to write. We might say that using professional writing tools leads to performance that is more like a professional writer. It is honest use of the real thing.

Vignette 3: Now, consider the following comment:

- TikTok is the MOST downloaded app on the Apple App Store. TikTok supports users in making videos, including videos that play in synchrony with other user videos. Video producers collaborate around the world to make duets, without ever meeting. Using TikTok is not about writing like a professional. TikTok is an entirely new medium, enabled by computation. It leads to writing and saying differently than one could without computation.

Vignette 4: Finally consider the following:

- Fortnite is one of the most successful video games of all time. In playing Fortnite, players use a broad range of computational tools to solve significant problems, from map navigation, to team collaboration, to managing complex ecological systems.

Few children get the opportunity to engage in these kinds of activities in their everyday world outside of the computer. The computational environment allows students to engage with complex and interesting problems. We can ask if these are honest versions of the problems, if students have deep understanding of what they are doing, and if they are developing skills for the real world—and we should ask those questions.

How do we prepare our children for never-seen-before problems?

The activity in Vignette 1 aligns with the notion that computational thinking is embodied in computer programming. Vignette 2 shows us it is not just programming that can impact thinking. A wide variety of computational activities can impact thinking. In Vignettes 3 and 4, we argue these activities illustrate "computational thinking"—though the activities in those vignettes have nothing to do with computer programming.

The users in those Vignettes are using computational tools to do computational thinking. They are using abstraction and decomposing problems, though they may not use those words. Much of the effort to implement computational thinking in schools has been about identifying the computing ideas and practices. Maybe the kids are already learning those, but on different terms, without our language.

People of the so-called baby-boomer generation may feel computational thinking is something special—and for them, it may well be. For the children growing up today, who are increasingly using digital tools to mediate their everyday lives, computational thinking is, well, just thinking! But that is just not enough. Learning to compute should give students a qualitative leap, so that they can think about new problems and think about the world in new ways.

How do we prepare our children for never-seen-before problems? We might start by redesigning TikTok and Fortnite.

Whether and Wither "Computational Thinking"

We already use computers to help many kinds of thinking, but much of that thinking would be the same without computers. We might get expanded thinking if we follow along the lines of extending mathematics and systems organizations to model complex situations that go beyond our commonsense reasoning, as seen in many scientific, engineering, medical, mathematical, and literary fields. Computing simulations has already revolutionized many fields. We might significantly impact society if all fields used this expanded thinking. So, there is a bird to be caught if we can sprinkle salt on its tail.

A strong rubric is "making systems about systems," and this accords well with the first ACM A.M Turing Award winner Alan Perlis' characterization of our field as "The science of processes; all processes." A subset of these processes are primarily algorithmic in nature, but to deal with the large range that computation can model, it is much more apt to "think all systems" and to see the representational possibilities of the computer make it a great fit to be the dynamic mathematics needed to make and understand systems.

This is a much larger—and in our opinion—much more useful characterization of computing as a subject in K–12, and it leads to a number of important differences from current practice. The big one is to help children learn about dynamic systems with interacting parts of all kinds, and how to make and model dynamic systems for deeper understanding (and considerable fun also!) Imagine something as engaging as Fortnite where the system is inspectable, where users might model their strategies and test them in simulation first, so that students might learn to use the power of expanded thinking.

A modeling and simulation point of view also serves to criticize the languages being taught today. For example, none of the common K–12 programming languages today are very good at modeling intercommunicating processes—despite both natural and human engineered systems working that way. Most of the languages that we put in school today can only handle one thread of control without ungraceful excursions into fragile and tricky designs.

We should not teach a qualitatively weak subset of something to children when we have better options. It might make later learning of a more powerful version more difficult.

Instead we should take our inspirations and goals from Jerome Bruner's assertion and challenge: "Any subject can be taught to any one at any age in an intellectually honest fashion if their level of development is heeded." Keeping the "intellectually honest" part means that—especially for young children—it will be necessary to invent real variants of adult versions of the subject matter—as has indeed been done so well by Montessori, Papert, Bruner, and others. We imagine a comprehensive suite of intellectually honest computing-based models for understanding systems can lead to much better notion of programming—for both adults and children. These will lead to much better programming language designs and environments as part of a larger curriculum made from the most powerful ideas about systems, processes, science, math, engineering, and computing itself.

One of the main ideas of K–12 schooling is to prepare children in general for their next phases of life, and subjects such as reading/writing/literature, science, mathematics, and history are taught to all to provide a "richness" of thought about both civilizations and how to be a citizen who supports civilization. Understanding civilization as a system is a powerful idea for all citizens. In our metaphor, we want citizens to participate in the redesign of the city and understand the rationale for its design. Students need fluency in order to be able to understand models and systems. Important thresholds of understanding must be reached before they can be part of one's thinking tools. Finding and inventing these thresholds for the general population of children, and how to teach to them, is the critical need of our time!

Representations to help thinking—language, mathematics, computing—are all best taught in context. Children should use computing with all the other fields of thought, rather than mostly in isolation. Rather than teach computer science as a separate topic that might transfer, we should teach with computational models in every field.

Conclusion: Montessori's Fortnite

We can and should improve schools to give students access to expanded thinking. We in computing have a powerful lever. We can change the computation.

Maria Montessori made the observation almost 100 years ago[3]: children are set up by their nature to learn their surrounding environment and culture. Changing the environment naturally leads to different learning. Montessori wanted her children to have qualitatively different thinking, so she invented new kinds of school.

Changing school today impacts only one part of today's children's lives. Changing computing impacts their environment both in and out of school. If Montessori were alive today, she would still want to redesign school, but she would likely want to change the computing, too. That is part of the child's whole environment. How would Montessori redesign Fortnite? What would she design instead of Fortnite?

Teaching computing as it is today is unlikely to have dramatic impact on students' everyday lives. It is our job to redesign computing, to give children new power to make sense of their world and change it. C

References

1. Clement, J. Algebra word problem solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education 13*, 1 (Jan. 1982), 16–30; doi:10.2307/748434
2. Dewey, J. *How We Think. A Restatement of the Relation of Reflective Thinking to the Educative Process* (Revised edition); D.C. Heath, Boston, MA, 1933.
3. Montessori, M. *The Montessori Method.* Frederick A. Stokes Company, New York, 1912.
4. Norris, C. and Soloway, E. Students write more, write better on the computer: Rigorously supported! *T.H.E. Journal*, (Nov. 11, 2017); https://bit.ly/2KQToTg
5. Wing, J.M. Computational thinking. *Commun. ACM 49*, 3 (Mar. 2006), 33–35; DOI: https://doi.org/10.1145/1118178.1118215

Mark Guzdial (mjguz@umich.edu) is Professor of Electrical Engineering and Computer Science, College of Engineering and Professor of Information, School of Information, University of Michigan, Ann Arbor, MI, USA.

Alan Kay (alan.viewpoints@yahoo.com) is Adjunct Professor, Computer Science, University of California, Los Angeles, USA. He is the recipient of the 2003 ACM A.M. Turing Award.

Cathie Norris (cathie.norris@unt.edu) is Regents Professor, Learning Technologies, College of Information, University of North Texas, Denton, TX, USA.

Elliot Soloway (soloway@umich.edu) is Arthur F. Thurnau Professor, Computer Science and Engineering, College of Engineering, University of Michigan, Ann Arbor, MI, USA.

The lead author organized the effort; all other authors are in alphabetical order.

DOI:10.1145/3363183

George Varghese

Interview

An Interview with Leonard Kleinrock

The UCLA professor and networking pioneer reflects on his career in industry and academia.

LEONARD KLEINROCK, DEVELOPER of the mathematical theory behind packet switching, has the unique distinction of having supervised the transmission of the first message between two computers. As a doctoral student at MIT in the early 1960s, Kleinrock extended the mathematical discipline of queuing theory to networks, providing a mathematical description of packet switching, in which a data stream is packetized by breaking it into a sequence of fixed-length segments (packets). ACM Fellow Kleinrock has received many awards for his work, including the National Medal of Science, the highest honor for achievement in science bestowed by a U.S. president.

UCLA Professor and ACM Fellow George Varghese conducted a wide-ranging interview of Kleinrock, an edited version of which appears here.

Leonard Kleinrock.

PHOTO BY DAN ANDERSON, COURTESY OF ELON UNIVERSITY/FLICKR (CC BY-ND 2.0)

GEORGE VARGHESE: Do you remember any epiphany as a boy that led you toward communication?

LEONARD KLEINROCK: I remember early in elementary school reading a *Superman* comic whose centerfold showed how to build a crystal radio out of household items that one could find on the street: a razor blade, some pencil lead, a toilet paper roll, and an earphone (which I stole from the telephone booth in the candy store down the street). I also needed a variable capacitor and had no clue what that was, but my mother, bless her heart, took me to a store in the electronics section of New York City, namely, Canal Street. The clerk helped me select the right part. Oh, the magic of listening to music from my newly built radio; and it required no battery or power at all. After that, I kept cannibalizing old radios and used the parts to design new radios that I put together. My mother never got in my way and allowed me a place behind our sofa to make a mess and to do my tinkering.

Your unusual college story should inspire some *Communications* readers.

My father fell ill and could not continue to run his grocery store, so I realized I could not attend college in day session and had no choice but to go to night school and bring home a salary by working full time during the day. That was a big blow. My father took me to an electronics firm where I could get a job serving as an electronics technician and eventually as an assistant electrical engineer doing industrial electronics. So instead of attending CCNY (City College of New York) as a daytime student, I attended at night. My day work was, however, wonderfully interesting: we were involved in designing and using photoelectric devices in many applications.

The people in night school were an interesting bunch—after all, who attends night school: crazies, dropouts, motivated students who had to work during the day, and GIs coming back from World War II (this was 1951) who were disciplined and very determined. The professors at night school worked in industry during the day so they had insight into practical matters. I remember a professor bringing a germanium transistor he worked on during the day to class saying "this is a better thermometer than an amplifier," and began to discuss ways to eliminate the temperature-dependent variations. This combination of combining practical issues with mathematical approaches has always supported my seeking to find intuition and insight behind theory. Claude Shannon, who was then—and still is—my role model, similarly had great insight and physical intuition into why things happened alongside his mathematical approach to problems.

You probably were thinking of getting a job after CCNY. How did you go to MIT instead?

I learned one day that an MIT professor was coming to CCNY at 4 P.M. to describe a terrific fellowship that would provide considerable financial support to pursue a master's at MIT as an MIT Lincoln Labs [a well-known R&D laboratory associated with MIT—*Ed.*] staff associate. I managed to get off work early that day, but when I asked the MIT professor for an application to the program, he told me they were available from a CCNY professor sitting at the back. The CCNY professor did not recognize me and when I told him I went to night school he said "get out of here." So I had to contact MIT directly to get a form. That I did and I was fortunate to be awarded the fellowship!

What was it like doing a master's at MIT as a Lincoln Labs associate?

My first supervisor at Lincoln was Ken Olsen, who later went on to found Digital Equipment and build the line of PDP computers. I worked in a group at Lincoln run by Wes Clark who built arguably one of the first PCs (the Linc computer). So there were a lot of brilliant people at Lincoln; and of course MIT professors would often visit.

What did you do your MS thesis on?

When I first got to MIT, I was interested in servomechanism systems and automatic control. Yet, my master's thesis at MIT was on optical readout of thin magnetic films for storage and processing. I made use of the Kerr magneto-optic effect whereby polarized light rotates differently when it reflects off a magnetized surface depending on the direction of magnetization. As a result, one could use polarized light to non-destructively "read" the bits on thin magnetic films (this was before disks). My job was to improve the reading process by amplification and coding. The thesis involved experiments and models and I even constructed a special digital logic using light bouncing off a sequence of thin films. My thesis must have impressed my MS supervisor—Frank Reintjes—at MIT because he insisted that I apply for a Ph.D.

But the idea was that after a Lincoln Labs fellowship you should work at Lincoln Labs as an engineer, right? And you had a first child coming by then?

That's right. Our first child was planned to come in the final summer just before I finished my MS, at which point I would be working full time at MIT Lincoln Lab and we needed the money since my wife would have to care for our newborn. So, I was not at all interested in pursuing a Ph.D. But Frank Reintjes was insistent and, amazingly, Lincoln Labs decided to offer me a follow-up Ph.D. fellowship to MIT just as they had done for my MS fellowship; this was a first for Lincoln. So I succumbed to the pressure and accepted the Ph.D. fellowship. Two others were also offered the Ph.D. fellowship: Larry Roberts [one of the founders of the Internet, see later—*Ed.*] and Ivan Sutherland [one of the founders of graphics and an ACM A.M. Turing Award recipient—*Ed.*] who both became lifelong friends.

Being surrounded by computers at MIT and at Lincoln Lab, it seemed inevitable to me they would eventually need to communicate with each other.

What were the first years of your MIT Ph.D. experience like?

Our Ph.D. qualifier was legendary for its difficulty with 50% of the applicants failing out like flies. My MS at MIT made it easier since the qualifying exam was largely based on the MIT MS curriculum, but full of trick questions. Interestingly Ivan (Sutherland) came in directly from Caltech (that is, without the benefit of exposure to the MIT MS material directly) and came out on top with one month to study; he is one heck of a smart guy. When I agreed to continue on with a Ph.D. program, I decided I wanted to work with the best professor I knew at MIT, and so called up Claude Shannon (founder of information theory). He surprised me (and shocked my friends) by inviting me to his house in Winchester, MA, USA. I remember the scene looking out on Mystic Lake as an automatic lawn mower (rigged up by Shannon) mowed the grass and his son's swinging hammock narrowly missed my head. Shannon wanted me to work on a strategy for the middle game in chess as part of a project that he and [AI Founder and Turing Award winner—*Ed.*] John McCarthy were working on.

How did you gravitate to what is considered your seminal thesis on packet communication?

I was looking for a fresh field to work on. It seemed to me that even Shannon had stopped working on information theory and coding theory, but most of my EE graduate student classmates wanted to work on it (with other professors). It seemed to me that the problems that remained seemed harder and less impactful. And chess was not my forte. At the same time, being surrounded by computers at MIT and at Lincoln Lab, it seemed inevitable to me they would eventually need to communicate with each other and the existing telephone network was woefully inadequate for such a challenge. Shannon agreed to be on my committee and my advisor was Ed Arthurs (by the way one other student of Arthurs was Irwin Jacobs of Qualcomm fame). Arthurs mentioned a classified project he had encountered for a network between computers. Here was an unmined area, an important area, one whose solution would have impact and one for which I had an approach—this was for me! I recognized that data communications generated highly bursty traffic and that the existing telephone network, which used the static assignment technology of (slow) circuit switching, was not up to the job.

I saw that what was needed was to assign (communication) resources in a highly dynamic, demand-based fashion, that is, dynamic resource sharing, wherein a resource is only allocated to a demand request when that demand needs it, and to then to release that resource when the demand no longer needs it. This concept is manifest today in so many systems (for example, Uber, AirBnB, seats on an airplane, and so forth) in what we often refer to as a shared economy.

My thesis proposal was entitled "Information Flow in Large Communication Nets." I was motivated by Shannon's teachings that large systems were especially interesting since, as systems scale up, emergent properties manifest themselves.

MIT was (and still is) a place of great intellectual ferment in those days. Tell us any memories you have of those days.

I remember the amazing collection of classmates that shared office space with me. For example, Jacob Ziv (information theory pioneer and inventor of Lempel-Ziv coding) and Tom Kailath (control theory pioneer) were part of the same suite with me. This lab housed Shannon's students along with others. I remember the stimulating conversations in which we engaged and taught each other our very different fields (information theory, control theory, networking, and so forth) and spurred each other on.

I had loved the few courses I taught while at MIT.

Queueing theory is widely used today but you may have been the first to apply and develop this tool for computer networks. How did that happen?

Queueing theory had been invented by the telephone engineers (starting with A.K. Erlang) in the early 1900s, then taken up by the mathematicians, but after the war the Operations Research folks began to apply it to industrial problems (for example, Jackson applied it to job-shop scheduling); but it never was a mainstream tool. Yet queueing systems models had all the ingredients of a mathematical approach to performance evaluation of networks that I needed since it dealt with ways of analyzing throughput, response time, buffer size, efficiency, and so forth. Further, and importantly, it was a perfect mechanism for implementing dynamic resource sharing.

The queueing books that were available in those days were very theoretical. I tried to remedy that later by writing a two-volume textbook called *Queueing Systems*, which was queueing theory for engineers and contained the first description of the ARPANET technology and its mathematical theory that was published in a book.

Yes, your book led me to research on networking as an undergraduate in IIT Bombay. I am sure it's influenced many others because of its clarity and strong intuition. Let's move on to your thesis on the mathematical theory of "stochastic networks." What does that mean?

My thesis dealt with computers exchanging messages—whose size and inter-interval times were governed by a probability distribution—across a network of what we would call routers today. Networks of steady deterministic traffic flows had been studied (for example, Max-flow Min-cut theorems had just appeared) and one node systems with stochastic arrivals had been well studied (queueing theory). However, very little had been done on the combination of those two, and this led to stochastic networks. This was a very hard problem to solve analytically, and to this day, the exact analysis is still intractable. However, I was able to crack the problem by making an assumption that the stream of traffic entering each router queue was a stream of independent traffic (the "independence assumption"); I was able to show via simulation that network behavior was accurately predicted with this assumption.

You were inducted into the Internet Hall of Fame in the first year when it opened, along with (Vinton) Cerf, (Robert) Kahn, and others. Your nomination says Leonard Kleinrock pioneered "the mathematical theory of packet networks, the technology underpinning the Internet." So how did stochastic networks morph into the Internet?

Once I recognized that the key issue was how to support bursty traffic in a data network, it became clear that the mathematics needed to represent the network had to be based on stochastic networks; this meant that I needed to extend queueing theory to the environment of networks, hence stochastic networks. There were at least two other independent threads that were nearly concurrent. Paul Baran at RAND corporation was tackling the problem of how to design a network for the military that was resilient to attack and so he hit on the notion of breaking messages into packets and dynamically routing them around failures and had simulations to show the effectiveness of this routing. He also proposed distributed network topologies that provided protection against partial network damage. His work was classified and so I did not see his papers until my thesis was completed. His work was right on! Another important thread was from Donald Davies who was working independently at the National Physical Laboratory in England and who realized that the packet switching was good because, as we had articulated, data was bursty. He coined the word "packet" and pointed out that long packets were more likely to contain an error than were small packets; hence he suggested packets of approximately 128 bytes, which was later used in the ARPANET design. He promoted packetization as having the desirable advantage of allowing small messages to swoop past large messages; interestingly, I had shown the exact form of this trade-off mathematically in my dissertation years earlier. Morover, he recognized that once messages were packetized, then retransmission of packets rather than whole messages would reduce delays in overall transmission. Further, he noted that the ability to pipeline packets reduced latency through the network.

So in some ways, Paul, Donald, and you explored different facets of the benefits of packet switching. Paul focused on routing resiliency, Donald on packet-level error resiliency, and you on mathematical performance evaluation and optimization of packet-switching networks using stochastic models. Is that accurate?

That is a fair characterization. I would add that Paul and Donald were looking mainly at critical architectural issues whereas I was more focused on extracting the underlying principles and developing a mathematical theory of packet networks. Among the principles were: dynamic resource sharing is key in an environment of bursty demands; large shared resources supporting lots of traffic are far more efficient than small resources supporting less traffic; and distributed adaptive control is efficient, stable, robust, fault-tolerant, and it works.

While we think of the Internet today to send email and support social networks, the motivation then was to share the expensive computers ARPA was funding.

Your thesis also anticipates and analyzes other benefits of packet switching we rely on today that are complementary to those pointed out by Davies. These include techniques such as priority queueing (for example VoIP is queued before data packets in today's routers) and splitting packets for the same destination across multiple paths (called ECMP). Is there anything else you want to mention before we move on from your thesis?

Some other aspects include the effects of scaling. I showed for the first time that in terms of performance, a single link of capacity C is better than N links each of capacity C/N (this was an example of the second principle I mentioned previously). I investigated how to optimally design network topologies, which contributed to the field of network flows that Howie Frank and Ivan Frisch and others made major contributions to. I also investigated distributed adaptive routing control but I modeled that by having each router precompute an ordered sequence of favorable routes for each destination and use the first route that was not congested locally.

Interesting! That's different from dynamic routing in today's Internet where routers use Dijkstra's algorithm to compute shortest paths. However, that technique takes longer to respond to failure. Some networks today (for example, MPLS protection) use your idea for faster recomputation of (possibly less optimal) routes after failure.

It is true that I did not suggest a Dijkstra-type updating procedure dynamically based on networkwide shortest paths. I introduced a simpler, albeit less optimal, protocol where the dynamics of the network were reflected in which links of interest out of a node were idle, indicating that the route to the destination using a queued link was not desirable and that the uncongested links leading out of a node to that destination were currently better choices.

Your thesis defense was a remarkable event ...

Larry Roberts, Ivan Sutherland, and I were very close friends and did our thesis defenses at the same time because we had all heavily used the MIT Lincoln Lab TX2 computer as part of our Ph.D. research. The union of our three Ph.D. committees came out to Lincoln Lab to view our work including Claude Shannon, who was on each of our committees, Marvin Minsky, and Peter Elias. It was a big heyday and a bit stressful given the credentials of those committee members. The projects were very different; we were just all using the TX2. Ivan did this great work on Sketchpad, Larry did his on machine perception of three-dimensional objects, and I did mine on communication networks. The TX2 allowed me to run an enormous simulation to verify the accuracy of my mathematical approximations.

You submitted your thesis and clearly it was well liked at MIT since they suggested you publish it as a book. How did you end up at UCLA and not at Lincoln Labs?

Morally, after their fellowship I felt I should work for Lincoln. But they were remarkably generous and offered to have me look around the academic and industrial circuit to see what opportunities were there. I received some great offers of research positions: Bell Labs, Lincoln Labs, Hughes, and many more. And then there were academic offers, including the one from UCLA for a tenure-track position (at half the salary I would get at Lincoln). But I had loved the few courses I taught while at MIT and realized I could augment my salary by consulting. So with the West Coast weather, the Wild West appeal, and a university position, I drove my family all the way across the country. Lincoln Labs was extremely gracious and even said that I could come back if I did not like it at UCLA—but it has been 56 years and I am still here!

Fast-forward to the birth of the Internet in a UCLA office. On October 29, 1969, you and Charley Kline, one of the Ph.D. students on your software team, transmitted the first message between computers hundreds of miles apart. What was the backstory?

In my software team, besides Charley there was Steve Crocker who headed the software group, Vint Cerf, and Jon Postel, all UCLA graduate students at that time and subsequently Internet luminaries. The backstory starts with Ivan Sutherland who became head of IPTO for ARPA in 1964. Ivan visited UCLA in 1965 and suggested we network the three nearly identical IBM 7090s on campus. But the three administrators didn't want to share their computers, so that network was never implemented.

How did it finally happen?

Bob Taylor (who later led Xerox PARC) took over IPTO after Ivan. Bob was convinced that IPTO needed a computer network to link the sites he was supporting so that they could share each other's computers and applications. Bob convinced Larry Roberts to come to Washington in 1966 and head up this idea of deploying a computer network. While we think of the Internet today to send email and support social networks, the motivation then was to share the expensive computers that ARPA was funding at sites like Utah (for graphics), Stanford Research Institute (databases, Doug Engelbart was there). Larry was familiar with my networking research and publicly credits my thesis for giving him confidence to spend millions of dollars of ARPA money on this crazy idea. Larry was also well aware of Baran's work and that of Davies (who had even built a single-node packet switch) and incorporated their ideas in the ARPANET design.

How did Larry get everyone together to create the ARPANET, the precursor to our Internet?

In 1967, Larry brought a bunch of us together to help him specify what this network would look like and what performance characteristics it would have. We specified the network and created the spec and then Larry put it out for bid. In December 1968, BBN was granted the contract.

In September 1969, BBN delivered the first IMP to you at UCLA. Why UCLA and not SRI or Utah?

My role in this ARPANET project was performance evaluation, design, experimentation, and measurement. At UCLA we had specified the measurement software BBN later implemented in each switch. It was natural that we would be the first node so that we could begin to conduct experimentation and make measurements of what was going on.

The first message on the Internet was "Lo" which seems to have Biblical connotations that go along with the Creation Story. Was this deliberate?

Not at all. We were trying to send the text "Login" to login to the SRI host but there was a bug and the software crashed after sending "Lo." Of course, the bug was in SRI's software, not ours nor in the network itself!

How did we get from the first ARPANET to the Internet we know today?

The first host-to-host protocol was called NCP but soon it became clear that the ARPANET would shortly be just one of many networks in an evolving "internetwork of networks" where every network would have a network number. The need for a more advanced internetworking protocol became clear and this was Cerf and Kahn's great achievement of TCP/IP for which they justly were given the ACM A.M. Turing Award.

The rest of the story, the commissioning of the NSF backbone, the decision to transition to multiple commercial backbones who had to cooperate, and so forth, are all well known. We had no clear idea of how the Internet would be used, but we caught our first glimpse when Ray Tomlinson introduced email in 1972 and it very quickly took over a major portion of the traffic; that was when it became clear that a major use would be to support people-to-people communication. Put another way, we completely missed social networking as a major use of the Internet. Indeed, it has been the case over and over again that the Internet community has been surprised each time major new applications have exploded in use (for example, the World Wide Web, peer-to-peer file sharing, blogs, user-generated content, search engines, shopping engines, social networks). What we are good at predicting is the underlying infrastructure of the Internet (networking technology, IoT, wireless access, mobility, and so forth). One other aspect of today's Internet we did not foresee was the emergence of the dark side (in all its manifestations) that plagues us today.

One other aspect of today's Internet we did not foresee was the emergence of the dark side (in all its manifestations) that plagues us today.

While the Internet was gaining steam, you trained several generations of remarkable students whose Ph.D. theses and papers with you greatly influenced the Internet and analyses of time-shared systems. Tell us more ...

There is so much to tell, so let me provide a small sample only. My first student was Ed Coffman, who worked on some extensions to priority queueing and time sharing. Most of my students who followed concentrated on various performance analyses of aspects of the Internet as it emerged. For example, the early ARPANET did synchronous (periodic) routing updates but Gary Fultz's thesis analyzed the benefits of asynchronous updates, something we take for granted today. Mario Gerla provided optimal routing design and provided an effective protocol. Parviz Kermani's thesis introduced the idea of cut-through routing, that is, starting to forward a packet as soon as the router read the destination address, thereby reducing latency, which is pervasive today in Local Area Networks. Farouk Kamoun's thesis introduced and showed the enormous benefits of hierarchical routing, which we see in OSPF areas today. Simon Lam and Fouad Tobagi initiated the analysis and design of wireless networking and provided the early analysis of Slotted Aloha (Lam) and CSMA (Tobagi). And so on.

And the triumvirate: Gerla, Tobagi, Lam—all full professors at UCLA, Stanford, and Texas (Austin) respectively—and winners of major networking lifetime awards.

Mario and I worked on network design techniques that went beyond my thesis while Howie Frank and Ivan Frisch were concurrently working on different techniques at Network Analysis corporation. Mario went to work for Howie Frank but we later hired him back at UCLA. Simon's thesis came out of the satellite packet switching meetings we had where he was my right-hand man. Out of that came his dissertation on the analysis of the instability of Aloha that Abramson had created in Hawaii. Then ARPA started moving to packet radio on a metropolitan area networking basis. The application was to foot soldiers or possibly tanks moving across a battlefield; that led to the SURAN survival radio network, and the whole packet-radio project. Tobagi and I started studying CSMA—carrier sense multiple access—which eventually contributed to Ethernet.

Their theses led to a cottage industry in both network design (popular in the 1970s) and to media access schemes (which continue to be popular today because of 802.11 and WiFi).

That early work caught the attention of researchers and industry as we continued studying the behavior of networking at large. We were fortunate to be working on these problems at an early stage.

I fear the power of the Internet is being lost.

Let's wind up by asking you about the future. In the old days, they would say "Young man, go West." You did that literally (Los Angeles) and metaphorically (the Internet was the new frontier, the Wild West of communication). Is networking now merely boring infrastructure like plumbing. What advice do you have for young researchers?

I think there is an enormous amount of exciting work to be done in networking and distributed systems in general. For example, areas that are in need of innovation, research, and development include IoT, distributed ledgers, the introduction of biologically inspired principles to networking (and engineering in general), distributed intelligent systems, advanced network architectures, network security, and much more. The space is awash with great problems to dive into. In the case of distributed ledgers, the technology that underlies bitcoin, I am excited on the one hand, but concerned on the other hand. What concerns me is that billions of dollars were poured into blockchain technologies soon after its birth, thus distorting its path to proper maturity; this is because profit-seeking companies and speculators jumped on the bandwagon right away, which may lead to brittle designs. By contrast, we had 20 years without commercial interruptions in designing the Internet and those years of careful curation helped, I believe, to make the Internet more robust.

What about future applications?

As I said before, it is hard to predict. We missed the social networking revolution completely. So, in some sense, we have created an Internet that is destined to continually surprise us with new, exciting, and explosive applications; that is a good thing because it represents creative opportunities for future generations of researchers. I think it would be interesting to study retroactively why certain applications (for example, Twitter, Facebook) grew so popular while others failed so that we can try to be better at prediction. Perhaps a lens based on behavioral economics of the sort pioneered by (Nobel Prize winning economist) Daniel Kahneman and his colleague, Amos Tversky may help.

What concerns you about the Internet?

I am concerned about the Balkanization of the Internet as nation-states cut off and censor Internet traffic and corporations create closed enclaves. I fear the power of the Internet is being lost. I realize this is partly because of security concerns but am confident about advanced technology developments that can ameliorate these concerns.

You have had a remarkable career of academic excellence (network performance evaluation clearly begun with your thesis and your later work with your UCLA students), real-world impact (you were heavily involved in the evolution of the Internet including helping write the famous Gore proposal), and entrepreneurship (you have started several companies including Linkabit (with Jacobs and Viterbi, which later led to Qualcomm) and Nomadix (an early mobile wireless company). What advice do you have for ACM members?

First, think deeply about the results of your work. It is not enough to evaluate your ideas. You need to keep thinking about them to distill principles before moving on to the next big thing. Second, try to bounce ideas among brilliant people. I have had the fortune of doing so with folks like Shannon, Sutherland, and colleagues like Gerald Estrin at UCLA. Third, aspire like Shannon to combine physical intuition with mathematical analysis. While I have done a fair amount of mathematical work, I am an engineer at heart. My early building of a crystal radio remains a watershed event in my life. C

George Varghese (varghese@cs.ucla.edu) is Chancellor's Professor, Computer Science, at UCLA, Los Angeles, CA, USA.

DOI:10.1145/3318157

Selena Silva and Martin Kenney

Viewpoint Algorithms, Platforms, and Ethnic Bias

How computing platforms and algorithms can potentially either reinforce or identify and address ethnic biases.

ETHNIC AND OTHER biases are increasingly recognized as a problem that plagues software algorithms and datasets.[9,12] This is important because algorithms and digital platforms organize ever-greater areas of social, political, and economic life. Algorithms already sift through expanding datasets to provide credit ratings, serve personalized advertisements, match individuals on dating sites, flag unusual credit-card transactions, recommend news articles, determine mortgage qualification, predict the locations and perpetrators of future crimes, parse résumés, rank job candidates, assist in bail or probation proceedings, and perform a wide variety of other tasks. Digital platforms are comprised of algorithms executed in software. In performing these functions, as Lawrence Lessig observed, "code" functions like law in structuring human activity. Algorithms and online platforms are not neutral; they are built to frame and drive actions.[8]

Algorithmic "machines" are built with specific hypotheses about the relationship between persons and things. As techniques such as machine learning are more generally deployed, concerns are becoming more acute. For engineers and policymakers alike, understanding how and where bias can occur in algorithmic processes can help address it. Our contribution is the introduction of a visual model (see the accompanying figure) that extends previous research to locate where bias may occur in an algorithmic process.[6]

Without proper mitigation, preexisting societal bias will be embedded in the algorithms that make or structure real-world decisions.

Interrogating Bias in Algorithmic Decision-Making

Of course, social bias has been long recognized. Some attribute the introduction of bias into algorithms to the fact that software developers are not well versed in issues such as civil rights and fairness.[3] Others suggest it is far more deeply embedded in society and its expressions.[4] Inspired by value chain research, while our model cannot resolve bias; it provides a template for identifying and addressing the sources of bias—conscious or unconscious—that might infect algorithms. What is certain is that without proper mitigation, preexisting societal bias will be embedded in the algorithms that make or structure real-world decisions.

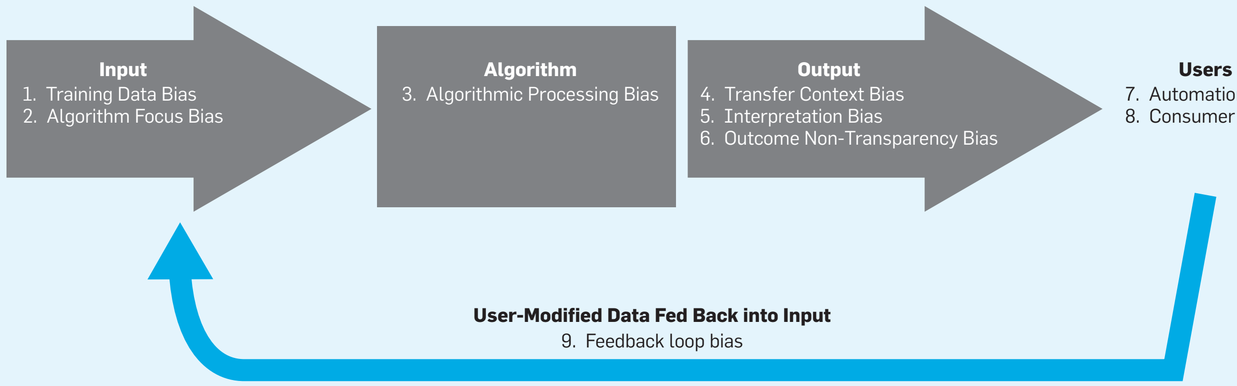
We model algorithm development, implementation, and use as having five distinct nodes—input, algorithmic operations, output, users, and feedback. Importantly, we incorporate users because their actions affect outcomes. As shown in the accompanying figure, we identify nine potential biases. They are not mutually exclusive, as it is possible for multiple, interacting biases to exist in a single algorithmic process.

Types of Bias

Training Data Bias. Predictive algorithms are trained on datasets, thus any biases in the training data will be reflected in the algorithm. In principle, this bias should be easy to detect, but the sources may be difficult to detect. Presumed gold standard datasets, such as government statistics or even judicial conviction rates, frequently contain bias. For example, if the criminal justice system is biased, then, absent corrections, the algorithm will mirror such bias. Thus, training sets can be subtle contributors to bias.

Algorithmic Focus Bias. Algorithmic focus bias occurs from both the inclusion and exclusion of particular variables. For instance, the exclusion of gender or race in a health diagnostic algorithm can lead to inaccurate or even harmful conclusions. However, the inclusion of gender, race, or even ZIP codes in a sentencing algorithm

Potential biases and where they may be introduced in the algorithmic value chain.



Source: The first six biases were adapted from Danks, D., & London, A.I. (2017).
The visualization and remaining materials are by Silva and Kenney.

can lead to discrimination. This is the conundrum: in certain cases, such variables must intentionally be used to produce less-biased outcomes.[5]

Algorithmic Processing Bias. Bias can be embedded in the algorithm itself. One source of such bias is the inclusion and weighting of particular variables. Consider the case of a firm's chief scientist's finding that "one solid predictor of strong coding is an affinity for a particular Japanese manga site."[10] If this is embodied in job-candidate-sorting software, then this seemingly innocuous choice might exclude particular qualified candidates. Effectively, a desired proxy trait inadvertently excludes certain groups that could perform the job.

Transfer Context Bias. Transfer context bias occurs when algorithmic output is applied to an inappropriate or unintended context. One example is using credit scores to make hiring decisions. Bad credit is equated with inferior future job performance, despite little evidence that credit scores are related to work performance. If the undesirable, but irrelevant trait is correlated with ethnicity, then it might lead to biased outcomes.

Interpretation Bias. Interpretation bias arises when users interpret algorithmic outputs according to their internalized biases. For example, a judge can receive an algorithmically generated recidivism prediction score and decide on the punishment or bail amount for the defendant. Because individual judges may be unconsciously biased, they may use the score as a "scientific" justification for a biased decision.

Outcome Non-Transparency Bias. Algorithms, particularly artificial intelligence and machine learning, often generate opaque results. The reasons for the results may even be inexplicable to the algorithm's creators or the software's owner. For example, when a machine-learning program recommends denial of a loan application, the bank official conveying the decision may not know the exact reasons for denial. The absence of transparency makes it difficult for the subjects of these decisions to identify discriminatory outcomes or even the reasons for the outcome.

Automation Bias. Automation bias results from the belief the output is fact, rather than a prediction with a confidence level. For instance, credit decisions are now fully automated and use group aggregates and personal credit history.[13] The algorithm gives certain people lower scores and limits their access to credit. Credit denial means their scores cannot improve. Often, the subjects and decision-makers are unaware of the algorithm's assumptions and uncritically accept the decisions. The European Union's GDPR's Article 22 has attempted to provide some protection by limiting automated algorithmic decision processes for legal or the equivalent life-affecting decisions.[11]

Consumer Bias. The biases that human beings act upon in everyday life are expressed in their online activities. Further, digital platforms can exacerbate or give expression to latent bias in online behavior. Users may consciously or unconsciously discriminate on the basis of a user profile that contains ethnically identifiable characteristics. Consumer bias can occur from either side, or party, in a digital interaction. Or, even more deliberately, anonymous online hackers purposely "taught" Microsoft's Tay chatbot, which was opened to the public for only a few days in 2016, to respond with racially objectionable statements. Effectively, the algorithm or platform provides users with a new venue within which to express their biases.

Feedback Loop Bias. Algorithmic systems create a data trail. For example, the Google Search algorithm responds to and records a query that becomes customized input for subsequent searches. The algorithm learns from user behavior. For example, in predictive policing, the algorithm relies almost entirely on historical crime data. Suppose the algorithm sends police officers into a neighborhood to prevent crime. Not surprisingly, increased police presence leads to higher crime detection, thereby raising the statistical crime rate. This can motivate the dispatch of more police, who make more arrests, thereby initiating a feedback loop. In another example, Google Search can learn that ethnically biased websites are often selected and therefore recommend them more often, thereby propagating them. As smart as algorithms can be, human monitoring continues to be necessary.

Benefits of Platforms and Algorithms

The potential benefits of algorithmic decision-making are less noticed, but it can also be used to decrease social bias. It is well known that members of the law enforcement community make decisions that are affected by a defendant's "demeanor," dress, and other characteristics that may correlate with ethnicity—an algorithmic process does not "see" these characteristics. This offers the potential for mitigating such bias. For example, Kleinberg et al. created a machine-learning algorithm that could do a better job than judges in making bail decisions.[7] The algorithm was optimized to reduce ethnic disparities among those incarcerated while also reducing the rate of reoffending. This optimization was possible because a disproportionately high number of people in certain racial groups are incarcerated. The point is that it is possible to design algorithms with different social goals. Critics ignore the fact the data and tools can be used to decrease inequity and improve efficiency and effectiveness.

Because algorithms are machines, they can be redesigned to improve outcomes. To illustrate, sales websites could reengineer a site to, for example, provide greater anonymity and thus reduce opportunities for consumer bias. Because all digital activities leave records, it is easier to detect biased behavior and thus reduce it. For example, a government agency could study online behavioral patterns to identify biased behavior. If it can be identified, then it can be prevented. For example, it would be easy to assess whether consumers are biased in their evaluations of online vendors and impose a standardization algorithm to mitigate such bias. Thus, while platforms and algorithms can be used in a discriminatory manner, they also can be studied to expose and address bias. Of course, the will to do so is necessary.

Conclusion

Computer scientists have a unique challenge and opportunity to use their skills to address the serious social problem of bias. We contribute to increased awareness by developing a readily understandable visual model for identifying where bias might emerge in the complex interaction between algorithms and humans. While we focus on ethnic bias, it is possible to extend our model to other types of bias. The model can be particularly useful in policy discussions to explain to policymakers and laypersons where a particular initiative could have an impact and what would not be addressed.

Interest in mitigating algorithmic bias has increased, but "correcting" the data to increase fairness can be hampered by determining what is "fair."

Interest in mitigating algorithmic bias has increased, but "correcting" the data to increase fairness can be hampered by determining what is "fair." Some have suggested that transparency would provide protection against bias and other socially undesirable outcomes.[2] Leading computing professional organizations such as ACM are aware of the problems and have established principles to guide their members in addressing these issues. For example, in 2017 the ACM Public Policy Council issued a statement of general principles regarding algorithmic transparency and accountability that identified potential bias as a serious issue.[1] Unsurprisingly, firms resist transparency, maintaining that revelation of their data and algorithms could allow other actors to game their systems. In many cases, this response is valid, yet it is also self-serving as it prevents scrutiny. Software developers often cannot provide definitive explanations of complex algorithmic outcomes, meaning transparency alone may be unable to provide accountability. Further, a single algorithmic model may contain multiple sources of bias that interact, creating greater difficulty in tracing its source. However, even in such cases, outcomes can be tested to discover evidence of potential bias.

Platforms, algorithms, software, data-driven decision-making, and machine learning are shaping choices, alternatives, and outcomes. It is vital to understand where and how social ills such as bias can be expressed and reinforced by digital technologies. Algorithmic bias can be addressed and, for this reason, critics who suggest these technologies necessarily will exacerbate bias are too pessimistic. Digital processes create a record that can be examined and analyzed with software tools. In the analog world, ethnic or other kinds of discrimination were difficult and expensive to study and identify. In the digital world, the data captured is often permanent and can be analyzed with existing techniques. Although digital technologies have the potential to reinforce old biases with new tools, they can also help identify and monitor progress in addressing ethnic bias.

References

1. ACM. Public Policy Council: Statement on Algorithmic Transparency and Accountability. (2017), 1–2; http://bit.ly/2n4RBjV
2. Ananny, M. and Crawford, K. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society 20*, 3 (Mar. 2018), 973–989.
3. Barocas, S. et al. Big Data, Data Science, and Civil Rights. arXiv preprint arXiv:1706.03102 (2017).
4. Caliskan, A., Bryson, J.J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science 356*, 6334 (2017), 183–186; https://doi.org/10.1126/science.aal4230
5. d'Alessandro, B., O'Neil, C., and LaGatta, T. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data 5*, 2 (Feb. 2017), 120–134.
6. Danks, D. and London, A.J. Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Aug. 2017), 4691–4697.
7. Kleinberg, J. et al. Human decisions and machine predictions. *Quarterly Journal of Economics 133*, 1 (Jan. 2017), 237–293.
8. Lessig, L. *Code: And Other Laws of Cyberspace* (2009); ReadHowYouWant.com.
9. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Broadway Books, New York, 2016.
10. Peck, P. They're watching you at work. *The Atlantic* (Dec. 2013); https://bit.ly/2jhKIt4
11. Portal, EU GDPR. Key Changes with the General Data Protection Regulation. EU GDPR Portal (2017).
12. Silva, S. and Kenney, M. Algorithms, platforms, and ethnic bias: An integrative essay. *Phylon: The Clark Atlanta University Review of Race and Culture 55*, 1–2 (2018).
13. Zarsky, T. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, and Human Values 41*, 1 (Jan. 2016), 118–132.

Selena Silva (ssssilva@ucdavis.edu) is a research assistant at the University of California, Davis, USA.

Martin Kenney (mfkenney@ucdavis.edu) is a Distinguished Professor in the Department of Human Ecology at the University of California, Davis, CA, USA, and is Research Director for the Berkeley Roundtable on the International Economy, Berkeley, CA, USA.

This research was funded in part by the Ewing Marion Kauffman Foundation and Clark Atlanta University. The contents of this Viewpoint are solely the responsibility of the authors.

India Region Special Section

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

wipro
TATA
₹

DOI:10.1145/3355552

Welcome to the India Region Special Section

WE ARE PLEASED to introduce the India Region special section to *Communications*' readers. The Indian subcontinent has a population close to 1.8 billion, and is unique due to its diversity of people, cultures, spoken languages, and wide disparities in socioeconomic conditions. The region plays an important role in the global computing landscape with its highly trained manpower, software companies, and top universities that produce students that not only serve local needs, but move around the world and have global impact. We developed this special section to mirror all these facets.

Last year, we publicized the plans for the special section and made an open call for contributions through ACM member channels and the ACM India website. We received 45 proposals through this period and augmented the list by reaching out to others to cover specific topics and areas. We also received proposals from colleagues in Sri Lanka and Pakistan. A selection of 22 outlines were identified for consideration. A workshop held at Microsoft Research in Bangalore on February 23, 2019 converged on the selection of 17 proposals to pursue as full articles. These articles underwent three rounds of reviews and modification. The final section presents nine articles covering Hot Topics and nine articles following Big Trends.

Computing solutions for the India region must deal with the scale of its population. We feature India's attempt at creating digital infrastructure and solutions at that scale, notably the biometric identification through Aadhar. The other big story is the success and reach of India's software industry; practically every piece of software sold in the world has components developed in India. The linguistic diversity of South Asia is a challenge when creating computer-based solutions, starting with suitable keyboards to the challenges of multilingual and mixed-lingual search. Another vibrant aspect of India is the resurgence of its start-ups. The articles in the special section attempt cover all these stories and more. The challenges of the socioeconomic milieu of this region are highlighted through articles on empowering women through mobiles, using speech to counter illiteracy, and in the problems faced by social media. The section also samples some of the research advances and challenges from this region.

We hope this collection of articles gives you a glimpse of the unique problems, opportunities, and exciting work in computing from the Indian subcontinent.

—P J Narayanan, Pankaj Jalote, and Anand Deshpande
India Region Special Section Co-Organizers

P J Narayanan (pjn@iiit.ac.in) is a professor and director of the International Institute of Information Technology Gachibowli, in Hyderabad, India. He is the former president of ACM India and former co-chair of ACM India Council.

Pankaj Jalote (jalote@iiitd.ac.in) is Distinguished Professor at the Indraprastha Institute of Information Technology Delhi, India, where he previously served as its founding director.

Anand Deshpande (anand@persistent.com) is the founder, chairman, and managing director of Persistent Systems, a global technology services company headquartered in Pune, Maharashtra, India.

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

INDIA REGION

SPECIAL SECTION

Hot Topics

Big Trends

IMAGES BY: (L) JOAT; (R) NOPPASIN WONGCHUM/SHUTTERSTOCK.COM

Machine Learning | DOI:10.1145/3355628

Extreme Classification

BY MANIK VARMA

WHAT WOULD YOU do if you had the superpower to accurately answer, in a few milliseconds, a multiple-choice question with a billion choices? Would you design the next generation of Web search engines, which could predict which of the billions of documents might be relevant to a given query? Would you build the next generation of retail recommender systems that have things delivered to your doorstep just as you need them? Or would you try and predict the next word about to be uttered by U.S. President Donald Trump?

Extreme classification has found applications in diverse areas ranging from information retrieval to recommender systems to computational advertising to natural language processing and even computer vision.

The objective in extreme classification, a new research area in machine learning, is to develop algorithms with such capabilities. The difficulty of the task can be judged from the fact that, even if it were to take you just a second to read out a choice, it would take you more than 30 years to go through a billion choices. In 2012, state-of-the-art multi-label classification algorithms were struggling to pick the correct subset of options in questions involving thousands of choices. Then, in 2013, a team from Microsoft Research India and IIT Delhi developed a classifier[1] that could scale to 10 million choices, thereby laying the foundations of the area. The approach was based on the realization that only a handful of choices would be relevant for any given question on average. The trick was therefore to quickly eliminate the millions of irrelevant choices. The classifier could then accurately and efficiently choose from the remaining hundred or so options.

Since 2013, extreme classification has come to be a thriving area of research in both academia and industry. Groups from Amazon, CMU, Columbia, Facebook, Fudan University, Google, Humboldt University, IIT Delhi, IIT Kanpur, Max Planck, Microsoft, MIT, Montreal, NEC, NUS, NYU, Stanford, Technion, TU Poznan, UC Davis, UT Austin, Yahoo, and others have developed a plethora of algorithms with varying trade-offs between the prediction accuracy, the prediction time, the training time of the classifier and its size. Most of these algorithms are either based on:

IMAGE BY R.CLASSEN

trees that learn a hierarchy over the space of choices so that approximately half the choices are eliminated at each node; embeddings that compress the number of choices by hashing them into a low-dimensional vector space; gating techniques that only consider the handful of relevant options for similar questions seen during training; and deep learning methods that learn the feature representation as well as the classifier. As a result, the community has made remarkable progress over the last six years with training times having been reduced by 10,000x, model sizes having reduced from terabytes to gigabytes, and prediction accuracies on benchmark tasks increasing from 19% in 2013 to 64% today. For instance, the Slice algorithm from MSR India and IIT Delhi, which won the best paper award at WSDM 2019, scales efficiently to problems involving 100-million choices and can be run on a laptop for small problems. Benchmark datasets as well as the source code for many of these algorithms are publicly available at The Extreme Classification Repository,[3] maintained by IIT Delhi and MSR India, which has become a vital resource for the community.

Extreme classification has found applications in diverse areas ranging from information retrieval to recommender systems to computational advertising to natural language processing and even computer vision. Many papers on extreme classification have been published in top-tier conferences in these areas including AAAI, AISTATS, CVPR, ECCV, IJCAI, ICML, KDD, NAACL-HLY, NeurIPS, SIGIR, WSDM, and WWW. Extreme classification has also opened a new paradigm for key industrial applications such as large-scale ranking and recommendation. For instance, extreme classification can be used to predict which of the top 100 million queries might lead to a click on a given ad or document. Similarly, extreme classification could also be used to predict which of the top 100 million videos you might wish to watch next. In certain cases, reformulating such problems as extreme classification tasks might increase revenue by millions of dollars as well as lead to performance improvements over traditional collaborative filtering, learning-to-rank, and content-based approaches. As a result, extreme classification has been deployed in various search and advertising products on the Microsoft Bing platform where it has significantly increased the ability of millions of people around the world to discover the products and services they are looking for. At the same time, extreme classifiers are also helping millions of small and medium enterprises by significantly increasing their sales and revenue, dramatically reducing the costs to reach relevant customers and enabling market growth by the discovery of new customers.

Extreme classifiers are helping millions of small and medium enterprises by significantly increasing their sales and revenue, dramatically reducing the costs to reach relevant customers and enabling market growth by the discovery of new customers.

Extreme classification has brought in many new research questions and technical challenges. A number of workshops have been organized at Dagstuhl,[2] ECML, ICML, NeurIPS, and WWW to discuss these questions. Watch the online videos from these workshops or check out The Extreme Classification Repository[3] if you are looking for an extreme challenge and want to help the community build the next generation of search engines and recommender systems.

References

1. Agrawal, R. et al. Multi-label learning with millions of labels: Recommending advertiser bid phrases for Web pages. In *Proceedings of the Intern. World Wide Web Conference* (Rio de Janeiro, Brazil, May 2013).
2. Bengio, S. et al. Extreme classification (Dagstuhl Seminar 18291). *v*, 7 (2019), 62–80.
3. The Extreme Classification Repository; https://bit.ly/2IDtQbS

Manik Varma (manik@microsoft.com) is a Principal Researcher at Microsoft Research India and an adjunct professor at the Indian Institute of Technology Delhi.

Women in Computing | DOI:10.1145/3355696

Designing ICT Interventions for Women in Pakistan

BY MARYAM MUSTAFA, AMNA BATOOL, AND AGHA ALI RAZA

INFORMATION AND COMMUNICATION technology (ICT) interventions are increasingly being used in developing countries to enable economic growth, employment, and empowerment. There is, however, growing agreement that the impact of ICTs in the Global South is not gender neutral but amplifies the existing gender inequalities within these countries.[2,7] This is also true for Pakistan and India, where most ICT interventions deployed have largely ignored the unique needs of the female Pakistani (48.63%) and Indian populations (48.53%). Multi-country research on the impact of ICTs reveals their great potential for bringing about positive socioeconomic change and gains in economic growth.[9] Similarly, studies reveal ICTs are one of the main drivers of economic growth in Asia, the Middle East, and Sub-Saharan African.[3,8] However, in order to ensure entire populations benefit from the deployment and adoption of ICTs an understanding of the specific needs and challenges faced by women is imperative.

Given the patriarchal structures that constrain women in Pakistan, traditional Western digital solutions do not work. Detailed ethnographies reveal specific cultural, religious, and social contexts ICT interventions must design around. We explore the specific needs and constraints of low-literate, low-income women in Pakistan and tackle the gendered design of technologies for financial inclusion, maternal healthcare, and digital social connectivity.

Financial inclusion refers to a process that ensures ease of access and affordability of financial services for a population and is an important means to tackle poverty and inequality. Financial exclusion is a crucial issue facing women in Pakistan, which is on the list of seven countries that constitute half of the unbanked population around the world where only 7%[a] of women are included in the formal financial sector (have access to financial services) compared to South Asia's average of 37% (in contrast, 76% of women in India own bank accounts[b]). In India, the main driving force behind increases in financial inclusion has been the 'Jan Dhan Yojana' scheme that mandated state-owned banks open at least one account for every unbanked household.

Given the patriarchal structures that constrain women in Pakistan, traditional Western digital solutions do not work.

Although digital financial services (DFS) are presented as a viable alternative to formal banking structures for the developing world, we show the concept of DFS, as it stands currently, is unsuited to the financial needs of Pakistani women. Unlike the developed world, mobile banking in Pakistan must account for phones as shared resources, secret/hidden financial transactions (women hide money from family/spouse), flexible and self-determined savings, and loan and dowry dynamics.[7] In the Western context, mobile phones are considered and designed for use as personal devices, unlike Pakistan where only 39% of women own mobile phones.[7]

Another crucial area where ICTs have great potential for impacting women's lives is healthcare. More than half a million women, 99% of them in developing countries, die each year from pregnancy-related causes, of which Pakistan alone is responsible for an estimated 30,000 deaths.[c] Large parts of Africa have tackled this using mobile phones to run services like MoTech and Abiye to deliver maternal and child health information.[5,6] Similarly, in Pakistan, we have launched mobile-phone-based information systems to give low-income mothers access to critical pregnancy information. Based on qualitative interviews with doctors and pregnant mothers, we designed and launched a combination of SMS text messages and interactive voice response (IVR) system that provides critical information for maternal care. Impact evaluations of the system reveal that targeted messaging has the largest, statistically significant impact on pregnancy-related knowledge.[4]

Another key issue faced by women in Pakistan is the lack of digital social connectivity. This is because all social media is based on the assumption of literacy and Pakistan is a country with an overall literacy rate of 58%; the literacy rate of women is 48%. To solve this, we have

[a] See http://bit.ly/2KE9Ma1
[b] See http://bit.ly/2F1omVE
[c] See http://bit.ly/2WAlEB5

launched voice-based community forums accessible over feature phones that function as social networks and allow women to anonymously create, enjoy, and share content.[1] Such social inclusion for women has allowed them greater connectivity and access to entertainment, education, and health opportunities.

Although our work has revealed essential guidelines for designing for women in patriarchal contexts some challenges remain. One open challenge is designing applications for shared mobile phones keeping in mind privacy for women in patriarchal contexts. Almost all applications, like mobile banking, email, Facebook, or Whats-App work on the assumption of a single user associated with a SIM. This is not true in Pakistan, where one phone/SIM is used by an entire family. How then, do you design and ensure privacy for each user? Similarly, given the harassment on voice-based social platforms that women face, how do you encourage female inclusion on these platforms?

A key issue faced by women in Pakistan is the lack of digital social connectivity. This is because all social media is based on the assumption of literacy and Pakistan is a country with an overall literacy rate of 58%; the literacy rate of women is 48%.

References

1. Ali, A. et al. Baang: A viral speech-based social platform for under-connected populations. In *Proceedings of the 2018 ACM SIGCHI Conf. Human Factors in Computing Systems* (Montréal, Canada, Apr. 21–26, 2018).
2. Arun, S., Heeks, R. and Morgan, S. ICT initiatives, Women and work in developing countries: Reinforcing or changing gender inequalities in South India? Institute for Development Policy and Management, University of Manchester, 2004.
3. Bahrini, R. and Qaffas, A.A. Impact of information and communication technology on economic growth: Evidence from developing countries. *Economies 7.1* (2019), 21.
4. Batool, A., Razaq, S., and Toyama, K. Maternal complications: Nuances in mobile interventions for maternal health in urban Pakistan. In *Proceedings of the 9th Intern. Conf. on Information and Communication Technologies and Development.* ACM, 2017.
5. Fajembola, T. Abiye: Safemotherhood: A case of leadership in turning the tide of maternal mortality in Nigeria. *Nigerian Health J.* (2011).
6. Mechael, P. MoTECH: mhealth Ethnography Report. Grameen Foundation, 2009.
7. Mustafa, M., et. al. Digital financial needs of micro-entrepreneur women in Pakistan: Is mobile money the answer? In *Proceedings of the 2019 ACM SIGCHI Conf. on Human Factors in Computing Systems* (Glasgow, Scotland, May 4–9, 2019).
8. Pradhan, R.P., Arvin, M.B. and Norman, N.R. The dynamics of information and communications technologies infrastructure, economic growth, and financial development: Evidence from Asian countries. *Technology in Society 42* (2015), 135–149.
9. Schreyer, P. The Contribution of Information and Communication Technology to Output Growth: A Study of the G7 Countries. OECD Science, Technology and Industry Working Papers 2000/2, OECD Publishing.

Maryam Mustafa (maryam.mustafa@cg.cs.tu-bs.de) is an assistant professor at the Lahore University of Management Sciences (LUMS), Lahore, Pakistan.

Amna Batool (amna.batool@itu.edu.pk) is a Teaching Fellow at Information Technology University, Lahore, Pakistan.

Agha Ali Raza (agha.ali.raza@itu.edu.pk) is an assistant professor at Information Technology University, Lahore, Pakistan.

PHOTO BY GARY YIM/SHUTTERSTOCK.COM

Query Optimization | DOI:10.1145/3347861

Turbocharging Database Query Processing and Testing

BY JAYANT R. HARITSA AND S. SUDARSHAN

DATABASE MANAGEMENT SYSTEMS (DBMS) constitute the backbone of today's information-rich society. A primary reason for the popularity of database systems is their support for *declarative* queries, typically in the SQL query language. In this programming paradigm, the user only specifies the end objectives, leaving it to the DBMS to automatically identify the optimal execution strategy to achieve these objectives. Declarative specification of queries is also central to parallel query execution in modern big data platforms.

Query processing and optimization have been extensively researched for close to five decades now, and are implemented in all contemporary database systems. Nevertheless, important challenges remain unsolved, and Indian universities have played a visible role in addressing these issues. As exemplars, we highlight recent research contributions on robust query processing, holistic optimization of database applications, and testing strategies for SQL queries and database engines.

Robust Query Processing

A crucial input to generating efficient query execution strategies, called *plans*, are the statistical estimates of the output data volumes for the algebraic predicates present in the query. In practice, these estimates, called *selectivities*, are often significantly in error with respect to the actual values subsequently encountered during query execution. The unfortunate outcome is a poor plan choice, resulting in query response times that may be worse by *orders of magnitude* relative to the optimal plan choice with the correct selectivities. A considerable body of literature exists on improving the statistical quality of selectivity estimates through sophisticated summary structures, feedback-based adjustments, and on-the-fly re-optimization of queries. However, a common limitation in this prior work is the inability to furnish performance *guarantees*.

A radically different approach that addresses the guarantee issue, called PlanBouquet,[a] has been recently developed at IISc Bangalore.[3] PlanBouquet completely *abandons* the classical estimation process for error-prone selectivities—instead, it employs a carefully calibrated "trial-and-error" sequence of time-budgeted plan executions that are progressively capable of handling more and more data until the query is eventually taken to completion. An advanced variant of this approach, called SpillBound, guarantees that the performance is *always* within a factor of (D^2+3D) relative to the ideal, where D is the number of predicates whose selectivity estimates may be erroneous.[5]

Further, empirical evaluations on industry-standard benchmarks have shown SpillBound to perform, in the worst-case, within a factor of 10–20 of the ideal, whereas contemporary database systems may suffer performance degradation factors running to the *1,000s* and beyond in such environments. This performance robustness of SpillBound is quantified in Figure 1 for a representative set of queries from the industry standard TPC-DS benchmark, the comparison yardstick being the PostgreSQL native optimizer.

These techniques represent an important milestone in the history of robust query processing since they are the first to provide quantitative performance guarantees, addressing a critical need of the database community.

Holistic Optimization

Database-backed applications often suffer from poor performance arising from sub-optimal ways in which imperatively written application programs access information from a database. For example, many application programs issue a long sequence of queries to a database, each of which requires a significant round-trip time due to latency in the database and network. Such inefficiencies cannot be addressed either by traditional database query optimizers or by traditional compiler optimizations. The DBridge system[b] developed at IIT Bombay therefore tackles this problem by rewriting application code to optimize data access.

DBridge carries out a set of potent equivalence-preserving transformations on imperative code

The DBridge suite of techniques brings the powerful benefits of declarative query optimization to imperative code.

[a] https://dsl.cds.iisc.ac.in/projects/QUEST

[b] https://www.cse.iitb.ac.in/infolab/dbridge

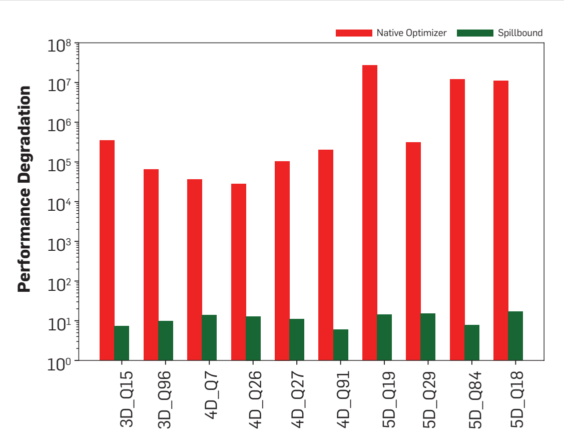


Figure 1. Performance robustness profile.

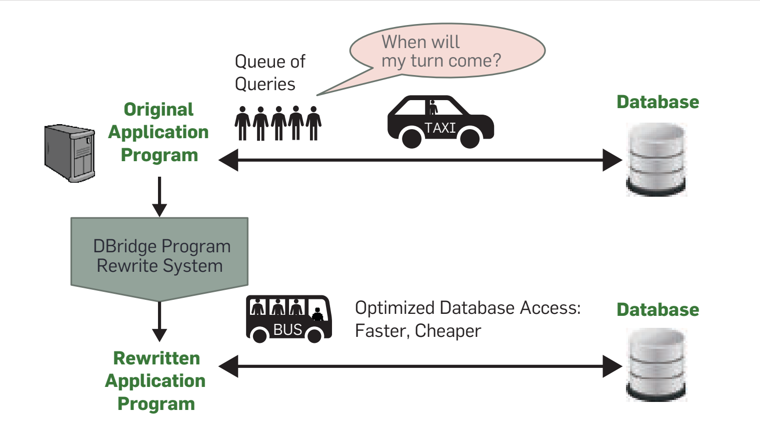

Queue of Queries
When will my turn come?
TAXI
Original Application Program
Database
DBridge Program Rewrite System
BUS
Optimized Database Access: Faster, Cheaper
Database
Rewritten Application Program

Figure 2. Rewrites for optimizing data access.

to speed up data access. The transformations successfully carry out batching and asynchronous submission of queries,[6] prefetching of query results, and conversion of procedural code to SQL. A metaphorical depiction of batching rewrites in DBridge is shown in Figure 2, where queries that are issued one-at-a-time, symbolized by the individual "taxis," are batched into a single unified request, carried by a "bus." Each transformation caters to a restricted scope and is therefore easy to prove correct, but in tandem they can successfully rewrite complex application programs. Further, the Cobra component of DBridge[4] efficiently chooses the least cost program from many alternative transformed programs, by leveraging concepts from query optimization based on algebraic equivalence rules.

Techniques for holistic optimization of queries containing imperatively coded user-defined functions (UDFs) were developed jointly by IIT Hyderabad and IIT Bombay; some of these mechanisms have subsequently been implemented and released in Microsoft SQL Server 2019, garnering excellent reviews from users.[c]

Collectively, the DBridge suite of techniques brings the powerful benefits of declarative query optimization to imperative code, opening a new research frontier. More details on these techniques may be found on the DBridge project home page.

Query and Engine Testing

With the onset of the Big Data world, where data is the engine driving virtually all aspects of human endeavor, it is vitally important to ensure both the applications and the underlying platforms are functionally correct. The XData system[d] developed at IIT Bombay supports testing of SQL queries by generating datasets designed to detect many types of common errors.[2] XData can be used in database courses to help students master the nuances of SQL query formulation and verify their correctness; further, the XData system facilitates *automated grading* of incorrect queries by assigning partial markings that reflect the severity of the errors. XData is currently operational at multiple universities.

The testing of Big Data platforms is addressed by the CODD project[e] at IISc Bangalore, using a distinctive metaphor of "dataless databases."[1] Here, databases with a desired set of characteristics can be efficiently *simulated* without explicit creation or persistent storage of the contents. This approach is essential since traditional testing techniques, which involve construction of representative databases and regression query suites, are completely impractical at Big Data scale due to the time and space overheads involved in their execution. The CODD tool has been successfully used for testing of database engines in the software and telecom industries.

Future Research

An important reason for the rapid adoption of SQL in the 1970s was its simplicity, which lent itself to effective query optimization. However, a host of complex features have been added over the years, and today's query processing world can be paraphrased as "with great expressive power comes great challenges." In this article, we have highlighted a few recent successes in tackling these challenges, but there remain rich opportunities for further contributions to the field. Productive future work areas include extending the holistic optimization concept to new domains (for example, machine learning), and leveraging query and data characteristics to deliver tighter robustness guarantees.

[c] https://www.microsoft.com/en-us/research/project/froid

[d] https://www.cse.iitb.ac.in/infolab/xdata

[e] https://www.cse.iitb.ac.in/infolab/xdata

References

1. Ashoke, S. and Haritsa, J. CODD: A dataless approach to big data testing. *PVLDB 8*, 12 (Aug. 2015), 2008–2011.
2. Chandra, B., Chawda, B., Kar, B., Reddy, K., Shah, S. and Sudarshan, S. Data generation for testing and grading SQL queries. *VLDB J. 24*, 6 (Dec. 2015), 731–755.
3. Dutt, A. and Haritsa, J. Plan bouquets: A fragrant approach to robust query processing. *ACM Trans. Database Syst. 41*, 2 (June 2016), 11–1:37.
4. Emani, K. V., and Sudarshan, S. Cobra: A framework for cost-based rewriting of database applications. In *Proceedings of the IEEE Intl. Conf. on Data Engg.* (Apr. 2018), 689–700.
5. Karthik, S., Haritsa, J., Kenkre, S., Pandit, V. and Krishnan, L. Platform-independent robust query processing. *IEEE Trans. Knowl. Data Eng. 31*, 1 (Jan. 2019), 17–31.
6. Ramachandra, K., Chavan, M., Guravannavar, R. and Sudarshan, S. Program transformations for asynchronous and batched query submission. *IEEE Trans. Knowl. Data Engg. 27*, 2 (Feb. 2015), 531–544.

Jayant R. Haritsa (haritsa@iisc.ac.in) is a professor at the Indian Institute of Science, Bangalore, India.

S. Sudarshan (sudarsha@cse.iitb.ac.in) is a professor at the Indian Institute of Technology, Bombay, India.

Digital Government | DOI:10.1145/3349629

Digital Transformation in the Indian Government

BY NEETA VERMA AND SAVITA DAWAR

IMAGE BY KONSTANTIN FARAKTINOV

DIGITAL INDIA is the flagship program of the Government of India with a vision to transform India into a digitally empowered society and knowledge economy. This program is centered on the vision of offering digital infrastructure as a core utility to every citizen, providing governance and services on demand, enabling the digital empowerment of citizens.[2] Besides policy making facilitation to the IT industry and start-ups, the government has also adopted state-of-the-art ICT for its own transformation for efficient and effective delivery of information and services to citizens at large. A specific focus has been on reaching the last mile as digital inclusion is at the core of the Digital India program. National Informatics Centre (NIC)[5] under the Ministry of Electronics and Information Technology is an important stakeholder in the digital transformation of the Indian government.

NIC is the driving force of the Digital India program and has also helped the government be in the forefront in the use of information technology. It has been working with the government for over four decades, providing state-of-the-art infrastructure, building solutions, as well as advising individual departments on action plans and adoption of appropriate technologies.

NIC set up the VSAT-based network for inter-government communication in 1982. The X400-based electronic messaging service was used by government long before the Internet was introduced to the country. With the help of NIC, the government has led the country in adopting the Internet and World Wide Web. The first Web presence of government was set up as early as 1995.

National infrastructure such as the network, cloud, video conferencing, government mail, GIS infrastructure, the public finance management system, and digital payments are key pieces which help provide a foundation for government departments to build IT systems that deliver services to citizens.

Centralized systems have had a huge impact on the work of government and the delivery of services.

Digital Platforms

Government has been using ICT-based systems to implement and manage its programs for over three decades. These systems have evolved with the advent of technology. Initially, client server systems were used, which had their own challenges as they had to be installed and maintained at the last mile. Over time, the government progressed to using Web-based systems and from there moved on to cloud-based systems. Although cloud-based systems provide a lot of advantages, good stable connectivity becomes a prerequisite for the success of any centralized system. The benefits that centralized systems provide are worth the investment done in provision of a stable and robust connectivity. With the proliferation of broadband and mobile telephony, this connectivity has significantly improved and many of the challenges overcome.

The centralized systems can help to create national-level data registries/databases, which offer great advantages to a huge country like India. The importance placed on data today can only be leveraged if these kinds of registries are developed and maintained centrally instead of as isolated silos. Such centralized systems optimize operations as they reduce maintenance costs and downtime. Furthermore, compliance to government regulations is easier and the integrated national registries help provide data from across the country while providing a single source for analytics. Centralized systems also address the many concerns over interoperability of various systems working at different levels and thus enhance delivery of services. Today, India has various national registries, such as the ones for driver's licenses, national vehicles, public distribution beneficiaries, and health registries.

The benefits of a national registry are seen in the eTransport project, which has successfully automated the Regional Transport Office across the country and set up a consolidated nationwide transport database with real-time updates and availability. A consolidated database of over 250 million vehicle records and over 150 million driver's license records already exists. With these registries, the transport department has evolved the eChallan application, wherein any police officer can issue an electronic citation or penalty on the spot, anywhere. These registries can also provide a close integration between vehicle insurance, pollution control systems, and accident reporting systems. A 360-degree profile of an individual or vehicle can be obtained. In the future, insurance premiums could be driven by such profiles.

Systems like these have helped bring about a digital transformation in India. These systems have had a huge impact on the work of government and the delivery of services. A description of changes in the public payment system and in the judiciary sector discussed here will serve as examples of their impact.

The centralized systems can help to create national-level data registries/databases, which offer great advantages to a huge country like India. The importance placed on data today can only be leveraged if these kinds of registries are developed and maintained centrally instead of as isolated silos.

Digital Transformations in the Financial Sector

The Public Finance Management System (PFMS)[6] has established itself as a safe, secure, efficient, and robust payment platform for the government of India. The system enables the successful delivery of payment from government treasuries and program agencies directly into beneficiaries' accounts.

PFMS was conceived as an online transaction system that not only helps the government manage its funds but at any point of time also provides a comprehensive view of the flow of funds across different wings of the government. Over time, the PFMS has built online interfaces with most of the banks in India. PFMS (as illustrated in Figure 1) is a very efficient and effective tool for monitoring of government funds.[7]

As part of the Digital India program, the government has leveraged on this unique position of PFMS and introduced the Direct Benefit Transfer (DBT)[1] of payments directly to the bank accounts of benefi-

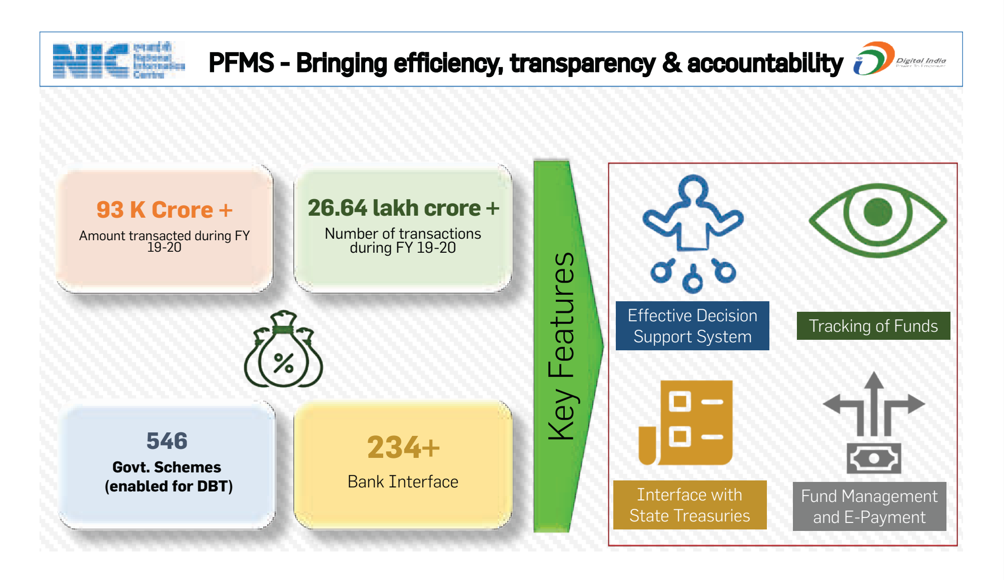

PFMS - Bringing efficiency, transparency & accountability
93 K Crore +
Amount transacted during FY 19-20
26.64 lakh crore +
Number of transactions during FY 19-20
%
546
Govt. Schemes (enabled for DBT)
234+
Bank Interface
Key Features
Effective Decision Support System
Tracking of Funds
Interface with State Treasuries
Fund Management and E-Payment

Figure 1. PFMS.

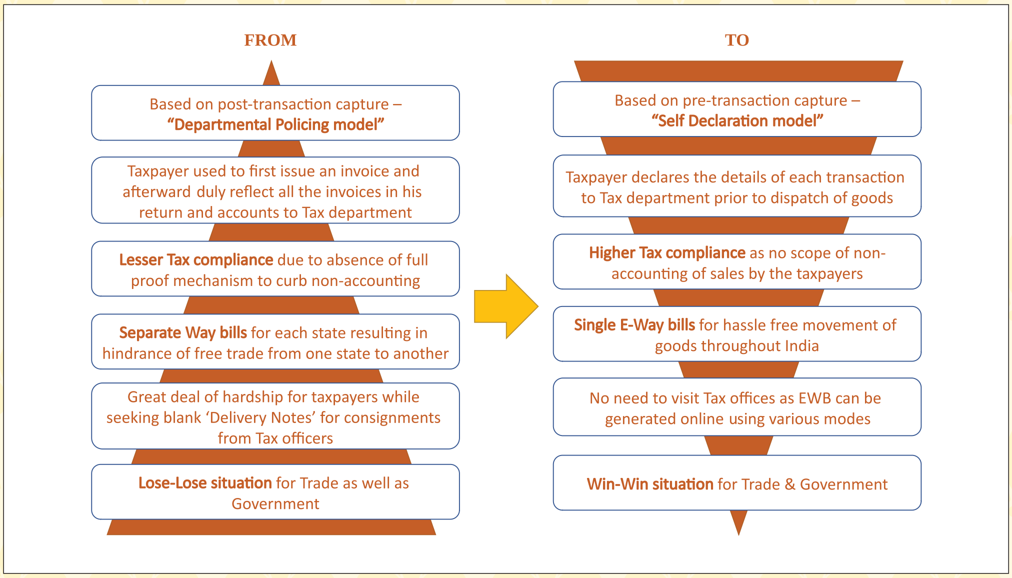

FROM
Based on post-transaction capture – "Departmental Policing model"
Taxpayer used to first issue an invoice and afterward duly reflect all the invoices in his return and accounts to Tax department
Lesser Tax compliance due to absence of full proof mechanism to curb non-accounting
Separate Way bills for each state resulting in hindrance of free trade from one state to another
Great deal of hardship for taxpayers while seeking blank 'Delivery Notes' for consignments from Tax officers
Lose-Lose situation for Trade as well as Government
TO
Based on pre-transaction capture – "Self Declaration model"
Taxpayer declares the details of each transaction to Tax department prior to dispatch of goods
Higher Tax compliance as no scope of non-accounting of sales by the taxpayers
Single E-Way bills for hassle free movement of goods throughout India
No need to visit Tax offices as EWB can be generated online using various modes
Win-Win situation for Trade & Government

Figure 2. e-WAY Bill.

ciaries. Various programs of the government that provide financial benefits or distribute subsidies to citizens have been integrated with PFMS. These programs cover social pensions, scholarships, employment guarantees, building of houses and toilets, and healthcare to name a few. In the past, funds traveled across various institutions or levels of government before reaching the beneficiary. With DBT, funds are now directly transferred into the beneficiary's bank account.

Electronic transfers have made a huge social impact as they ensure the timely transfer of benefits to citizens, bringing efficiency, effectiveness, transparency, and accountability to the system. Further, the government is able to ensure accurate targeting of beneficiaries and most importantly overcome other nuances of multi-layer transfer of funds, thereby eliminating pilfering and curbing leakage and duplication. DBT is further strengthened by the introduction of Aadhaar-based payments. Aadhaar is a 12-digit random number issued to residents of India by the Unique Identification Authority of India (UIDAI).[8] The Aadhaar-enabled Public Distribution System has helped ensure the availability of food to over 330 million poor people at affordable prices, thus enhancing their food security (see the article by Raghavan et al. on p. 76).

PFMS together with DBT has brought about phenomenal change in terms of social impact. State governments recognized this and are also leveraging the system to transfer benefits under their programs. An estimated 100 billionINR (US$1.43 billion) is the annual gain to NIC from the PFMS platform. Integration with treasuries and the linkage of Aadhaar and DBT has helped government save close to 830 billionINR (US$11.5 billion).

The Goods and Services Tax (GST) is an indirect tax levied on the supply of goods and services. It is a multistage, destination-based tax that is levied, for example, at every step as a product moves from materials through production then distribution and sale. When GST was introduced in July 2017, the e-Way Bill[3] was also introduced to allow a common permit for movement of goods throughout the country. e-Way Bill is an electronic document that includes details regarding the movement of goods; it must be carried by transporters for any consignment over a certain threshold. The e-Way Bill mechanism ensures goods are transported in accordance with GST laws and that taxes are paid for the supply of goods (see Figure 2).

Electronic transfers have made a huge social impact as they ensure the timely transfer of benefits to citizens, bringing efficiency, effectiveness, transparency, and accountability to the system.

Through the e-Way Bill, taxpayers, transporters, and tax officers all rely on a unified system. The implementation of the e-Way Bill has helped boost GST revenue collections, abolished post-dated checks, and increased tax compliance. There has been significant improvement in the ease of doing business due to the self-declaration and reporting enabled by e-Way Bills, which also save time in the transport of goods. Approximately 700,000 e-Way Bills are generated every day.

Revolutionizing the Indian Judiciary Sector

NIC is the single organization that consults and interacts with government at different tiers throughout India, from central government to state government to district administration. It is also the only organization that works across the three organs of state, namely the executive, judiciary, and legislative branches. The eCourts[4] ICT system is helping transform the Indian judiciary by enabling courts to enhance judicial productivity and provide citizen-centric services. The system, as illustrated in Figure 3, ensures service delivery and promotes transparency to all stakeholders, including litigants, advocates, judicial officials, and police officers. The eCourts project is implemented in more than 3,091 court complexes—the last mile courts scattered across the country. With over 627 districts online, data of a staggering 116.3 million cases and 91.5 million judgments and orders are available online.

The National Judicial

Data Grid (NJDG) has brought transparency to the country's justice delivery system. Tracking pending litigation at the district level has also opened judicial matters to the general public, researchers, academicians, and society at large.[a] NJDG also serves as a decision support system to authorities like the Supreme Court, high courts, the central government, and state government to monitor pendency on varied attributes for effective decision making.

Conclusion

NIC is the IT arm and an integral part of the Indian government. This single organization consults and interacts with government institutions at all tiers, from the central to panchayat (village) level. In addition, NIC has also set up nationwide infrastructure that is leveraged by all these institutions in their internal functioning as well as the delivery of services. This structure is unique in the world and has accelerated the adoption of new technologies by government at all levels. The ready availability of infrastructure like a government network, datacenters, the cloud, and mail has fast-tracked implementation of various initiatives under the Digital India program. Cybersecurity of these infrastructure systems is also managed by NIC, making it versatile and unique.

NIC can be considered the prime builder of e-government applications and services as well as a promoter of digital opportunities for sustainable development. Use of open source technologies and open standards is at the core of many of the projects implemented by NIC. This has reduced the reliance on proprietary software and enhanced interoperability. These governance and citizen-centric products have proved a great impetus to citizen empowerment and resulted in a vast transformation in the delivery of government services, wider transparency, decentralized planning and management, and better efficiency and accountability to the people of India.

NIC's role in e-governance initiatives is leading to a truly Digital India and ensuring effective citizen-centric governance. The imprints of NIC can be seen in almost every sector of the government such as health, education, transport, agriculture, to name just a few. With several such nationwide flagship initiatives and services, NIC is spearheading the country's growth in the digital realm and contributing to its inclusive development. There has been massive savings for the government in this digital transformation and direct financial benefits to citizens. Citizens are the ultimate winners, with quicker, transparent delivery of services and benefits.

NIC can be considered the prime builder of e-government applications and services as well as a promoter of digital opportunities for sustainable development. Use of open source technologies and open standards is at the core of many of the projects implemented by NIC.

Further, with the advent of Digital India, the huge amount of data generated through e-governance initiatives is being used for effective planning and decision making by the government, as NIC provides support in the domain of data quality assessment and big data analytics. Keeping pace with emerging technologies, NIC has started to incorporate technologies such as deep learning, linguistic analysis, and advanced analytics in its products and e-governance applications for greater societal benefits.

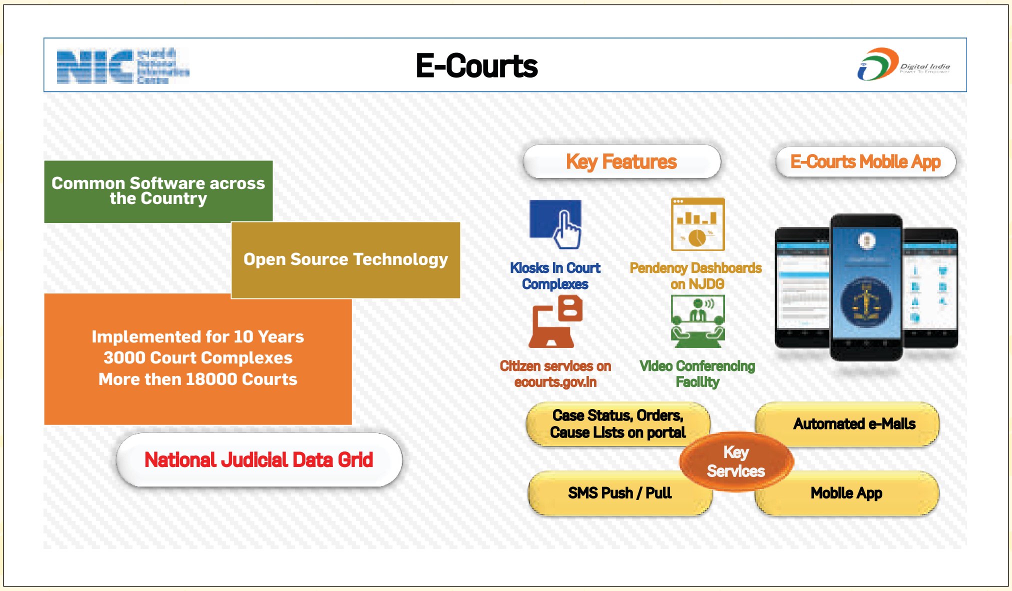
[a] http://njdg.ecourts.gov.in

References

1. DBT https://dbtbharat.gov.in/
2. Digital India https://digitalindia.gov.in/
3. E-Way bill https://ewaybill.nic.in/
4. Ecourts https://ecourts.gov.in/ecourts_home/
5. National Informatics Centre https://www.nic.in/
6. PFMS https://pfms.nic.in/NewDefaultHome.aspx
7. Sengupta, D.and Shastri, N. Digital Payments through PFMS—Facilitating digital inclusion and accelerating transformation to a 'Digital Economy.' In *Proceedings of the 12th Intern. Conf. Theory and Practice of Electronic Governance* (Apr. 2019).
8. Unique Identification Authority of India; https://uidai.gov.in/

Neeta Verma is Director General of the National Informatics Centre in New Delhi, India.

Savita Dawar is Deputy Director General of the National Informatics Centre in New Delhi, India.



E-Courts

Common Software across the Country

Open Source Technology

Implemented for 10 Years
3000 Court Complexes
More then 18000 Courts

National Judicial Data Grid

Key Features

Kiosks in Court Complexes

Pendency Dashboards on NJDG

Citizen services on ecourts.gov.in

Video Conferencing Facility

E-Courts Mobile App

Case Status, Orders, Cause Lists on portal

Automated e-Mails

Key Services

SMS Push / Pull

Mobile App

Figure 3. eCourts.

Education | DOI:10.1145/3343445

CSpathshala: Bringing Computational Thinking to Schools

BY VIPUL SHAH

BHUMIKA AND PUSHKAR, 12-year-old students from a government school in the village of Takalkarwadi, in Khed, Maharashtra, are playing the "Guess My Birthdate" game. The goal of the game is to find the date by asking the least number of questions. The students' strategy is to analyze each question in terms of the number of dates it eliminates.

"Guess my Birthdate" activity at Takalkarwadi School, Maharashtra.

Some 300,000 students from 750 schools in 11 states throughout India are learning computing through "unplugged" activities as part of CSpathshala,[1] ACM India's education initiative. The name CSpathshala is derived from computer science and Pathshala, which means place of learning or a school. Launched in 2016, CSpathshala's primary goals are to promote computer science education in K–12, to influence policymakers to introduce computing into mainstream curricula, and to train teachers so that every child in India learns computing as a science by 2030.

The National Policy on ICT for School Education in India[6] advocates the development of a model Curriculum for ICT that would include conceptual knowledge enhancement and enable the development of generic skills with focus on digital literacy. Although teaching computer science has already been introduced in urban India, it focuses primarily on digital literacy and a bit of programming.

Some 300,000 students from 750 schools in 11 states throughout India are learning computing through "unplugged" activities as part of CSpathshala, ACM India's education initiative.

Introducing a computing curriculum has not been easy and has posed several unique challenges:

- *Scale:* As per government reports,[7] India has over 1.6 million schools offering K–12 education to 300 million students. To compound the problem, in addition to two national boards of education, each of the 29 states in India has its own education board! While English is the common language of instruction in the urban areas, 70% of the population residing in the rural areas is educated in the state's regional language.
- *Infrastructure:* 63% of the schools have electricity and only 27% of schools have computers. In rural areas, electricity may be available for a few hours a day and the school may have only 1–2 computers. Urban schools are better equipped with computer labs that allow a computer to be shared by 2–3 students.
- *Teacher skills:* A survey we conducted corroborated findings in Raman et al.[12] Teachers from rural areas had no computing background. Moreover, only 59% of the teachers working in urban areas had exposure to some form of computing education, with only 10% having a computer science degree.

A national curriculum committee explored the CSTA K–12 curriculum framework and recommendations,[9] CAS U.K. curriculum,[2] code.org lessons,[a] Computer Masti,[8] and CS unplugged material,[4] and have developed an unplugged computing curriculum[5] influenced by the New Jersey discrete math curriculum for problem solving.[10] It includes topics like systematic listing, counting and reasoning (systematically arriving at all possible answers and reasoning on completeness), iterative patterns and processes (looking for patterns to generalize and apply to given problem), organizing and processing information (data collection, representation, and analysis), discrete mathematical modeling (abstractions like graphs and trees), following and devising instructions (initially following, then devising a precise set of instructions and later evaluating multiple solutions) and programming.

Strategies to address the challenges mentioned here include:

- Efforts have been directed toward carrying out a pilot program with 500+ rural government schools and working with 2–3 education boards. The Tamil Nadu state education board has adopted computational thinking as part of its math curriculum for 10,000 schools. Another state educational board will begin a pilot shortly with

[a] www.code.org

CSpathshala's primary goals are to promote computer science education in K–12, to influence policymakers to introduce computing into mainstream curricula, and to train teachers so that every child in India learns computing as a science by 2030.

1,500 schools. Teaching aids have been translated into three regional languages enabling reach beyond English medium schools.

- Developed an unplugged curriculum to overcome lack of infrastructure.
- Prioritized teacher training and creation of teaching aids. Some 250+ CSpathshala volunteers have created teaching aids for 200+ lessons for grades 1–8 that are distributed under CC license; 3,700 teachers from 1,850 schools have been trained through 70 training programs, all at no cost to schools.

From 5,000 students in 15 pilot schools in 2016–2017, the initiative has been steadily making inroads. Cambridge University Press has partnered with CSpathshala to publish CS educational books, thereby increasing the reach. While the feedback from teachers has been very encouraging, the annual conference on computational thinking for schools[11] revealed that an increasing number of teachers are integrating computational thinking with math curriculum as well as developing innovative pedagogical methods to engage students. Teachers are applying a systematic problem-solving approach and extending it to other subjects. Formal studies will be undertaken to measure the impact of the program.

With CSpathshala, a formal computing education is now available to students in rural India who have traditionally been deprived of the same.

References

1. ACM India's education initiative CSpathshala: Bringing computational thinking to schools in India; www.cspathshala.org
2. CAS-UK. Computing at School Working Group http://www.computingatschool.org.uk
3. Computer Science Teachers Association. https://www.csteachers.org/
4. Computer Science Unplugged: csunplugged.org/
5. CSpathshala curriculum; https://cspathshala.org/curriculum/
6. Department of School Education and Literacy Ministry of Human Resource Development Government of India. *National Policy on Information and Communication Technology In School Education 2012*; http://bit.ly/2K5ULyt
7. Government's Unified District Information System for Education 2016 Report; http://bit.ly/31neqPO
8. Iyer, S., Khan, F., Murthy, S., Chitta, V., Baru, M. and Vishwanathan,U. *CMC: A Model Computer Science Curriculum for K-12 Schools*, 2013.
9. K–12 Computer Science Framework. https://k12cs.org/
10. New Jersey Mathematics Curriculum Framework, 1997; http://bit.ly/31qEg5w
11. *Proceedings of the First Conference on Computational Thinking in Schools.* (Pune, India, Apr. 2019).
12. Raman, R., Venkatasubramanian, S., Achuthan, K. and Nedungadi, P. Computer science education in Indian schools: Situation analysis using Darmstadt model. *Trans. Comput. Educ. 15*, 2, Article 7 (May 2015)

Vipul Shah (v.shah@tcs.com) is Principal Scientist at Tata Consultancy Services, Pune, India.

The CSpathshala education initiative teaches computer science using "unplugged" activities.

ALL PHOTOS COURTESY OF CSPATHSHALA.ORG

Finance | DOI:10.1145/3343458

Creative Disruption in Fintech from Sri Lanka

BY AJIT SAMARANAYAKE, SAMPATH TILAKUMARA, THAYAPARAN SRIPAVAN, AND RASIKA WITHANAWASAM

DURING THE 1990S, the Sri Lankan IT sector was sandwiched between the forces of free market competition and the internal turbulence due to civil unrest. The relatively small internal marketplace made it difficult to attract foreign investments and expand businesses beyond IT offshoring. However, stock trading was a brick-and-mortar business that presented promising growth potential with the advent of financial technology (fintech).

Figure 1. The Colombo Stock Exchange (CSE), the first customer. (Photo courtesy of CSE)

Sensing this opportunity, a set of seasoned managers at an existing IT services business set off with a broader vision, marking the birth of MillenniumIT (now known as LSEG Technology), a fintech product company. The farsighted entry into fintech, and the experience in mobilizing local talent, contributed to the early success of MillenniumIT.

MillenniumIT introduced novel concepts in designing complex electronic trading systems with predictable performance that met the regulated ultra-high resiliency requirements. Being an early mover allowed for the slow, steady penetration of the company's technology into capital markets around the world. In 2009, the London Stock Exchange Group acquired MillenniumIT and secured the status of the "fastest trading system" through a subsequent technology upgrade. Today, LSEG Technology[6] (as it is now known) is a key contributor to the overall group's technology strength, and powers over 40 capital market institutions around the globe.

The early application of 'creative disruption' to a niche market with immense growth potential has proven to be a very effective tool and strategy.

Since its inception, creative disruption along four major themes has served to keep LSEG Technology competitive.

1. Software resilience for scalable distributed systems.

Scalable distributed systems are at the core of the architecture of the organization's electronic trading systems. Competitive challenges served to help contain the costs of scalability, while preserving high resiliency. LSEG Technology quit using costlier hardware-dependent resiliency by introducing software fault-tolerance models into a freshly built common technology framework for fintech applications.

Such software fault tolerance introduced patented models of replication, synchronization, and recovery that ran on commodity hardware at a fraction of the cost, establishing new industry benchmarks on system availability for mission-critical fintech applications.

A first-of-its-kind deployment in Sri Lanka's national stock exchange (CSE) in 1995 was followed by implementations in global trading hubs in London, Milan, Oslo, and Johannesburg.

2. and 3. High-performance and heterogeneous computing.

Execution of complex functionality at ultra-low latency is imperative for electronic trading systems.

That pattern of confounding expectations led to a number of inflection points when LSEG Technology offered an ultra-low-latency trading system to address the London Stock Exchange's requirements in 2011. The low-latency external interfaces developed as part of this solution allowed co-located high-frequency traders to take advantage of the ultra-low latency of the platform. A sub 100µs step-jump in end-to-end latency was possible with a full-stack re-architecture, and being the first to infuse emerging transport technologies (such as Infiniband) helped the London Stock Exchange gain market share and stay ahead of other leading exchanges.

In 2014, an award-winning[8,9] low-latency market data distribution platform was introduced with the use of field programmable gate arrays (FPGA)[2] that yielded a 95% performance improvement (sub 5µs end to end) compared to homogeneous software.[3–5,7] The latest generation of the heterogeneous (FPGA, GPU) application suite enables new business models by realizing financial risk simulations and deep learning in real time.

4. Description-driven systems.

The description-driven approach to software generation was again a disruptive response to meet demands for higher quality and quicker delivery, with lower costs. LSEG Technology introduced a patented business rule engine in 1998, which allowed flexibility in specifying business features without requiring redeployment or upgrades. The core of this approach was extensible to the description of an entire system (that is, data model, business functionality, work flows, user interface (UI), and deployment). Using a combination of theorem provers and code generators, it has been shown to generate approximately 80%–90% of the code of a system. An initial prototype of a fully functional post-trade clearing system demonstrated that a 5k-line system description will generate up to 950k lines of C++, JavaScript, or SQL code.

Summary

The success of LSEG Technology bears testimony that despite contextual barriers, organizations in this region can indeed become globally competitive technology leaders in specialized niche markets. The early application of 'creative disruption' to a niche market with immense growth potential has proven to be a very effective tool and strategy.

The increasing pace of technological advancements warrant a balanced outlook toward agility, far-sighted bets on technology, and investments in intellectual capital, to exploit the unfolding opportunities of the future.[1]

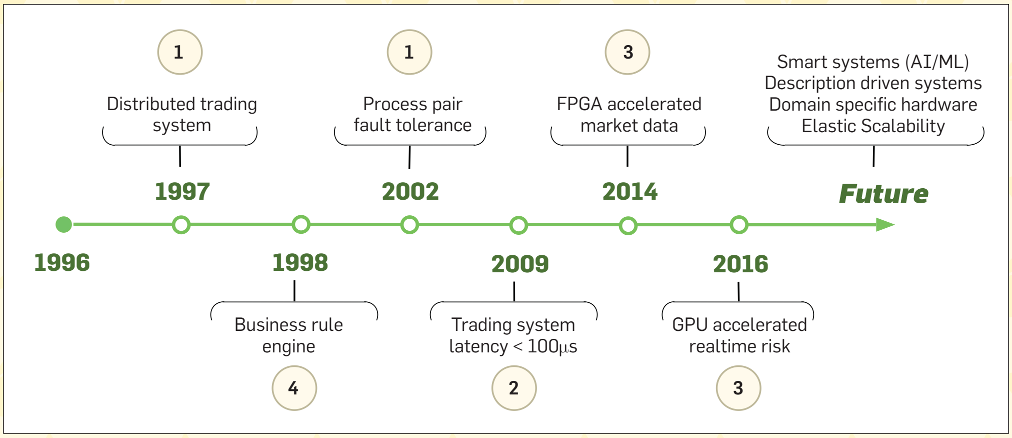

1
Distributed trading system
1997
1
Process pair fault tolerance
2002
3
FPGA accelerated market data
2014
Smart systems (AI/ML)
Description driven systems
Domain specific hardware
Elastic Scalability
Future
1996
1998
Business rule engine
4
2009
Trading system latency < 100µs
2
2016
GPU accelerated realtime risk
3

Figure 2. Themed milestones of creative disruption.

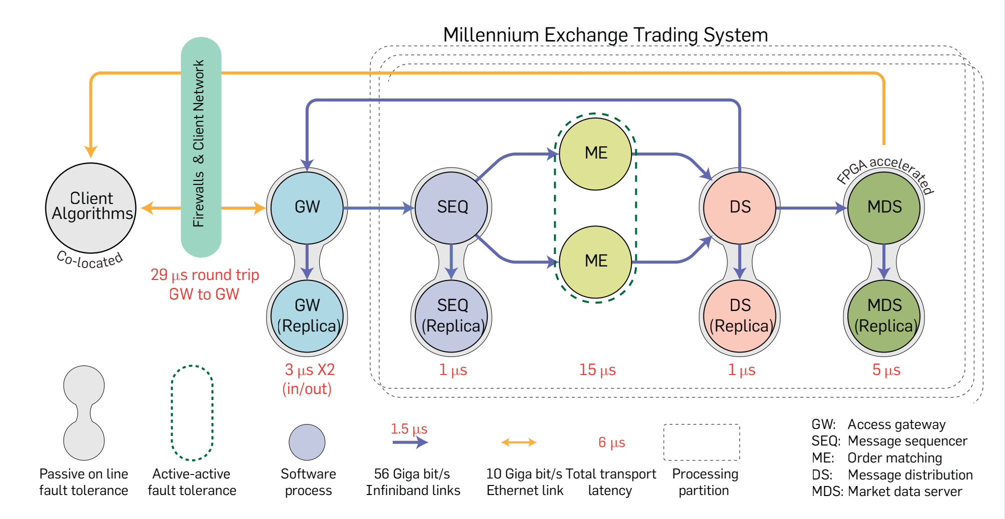


Figure 3. An abstract view of the Millennium Exchange trading system.

References

1. Bloomberg LP U.S., 2019; https://bloom.bg/2KG8NGl
2. Businesswire.com. 2013; http://bit.ly/2R5Brlm
3. Finextra.com. 2016; http://bit.ly/2wMtF6D
4. Fnlondon.com. 2016; http://bit.ly/2K79sRC
5. Ibsintelligence.com. 2016; http://bit.ly/2WyHC7w
6. London Stock Exchange Group PLC, U.K. 2019; https://www.lseg.com/lseg-technology
7. Thomsonreuters.com. 2015; https://tmsnrt.rs/2wNWvUg
8. Waterstechnology.com. 2017; http://bit.ly/31pIsSV
9. Waterstechnology.com. 2018; http://bit.ly/2wOLvpM

Ajit Samaranayake (ajit@lseg.com) is chief scientist at LSEG Technology, Colombo, Sri Lanka.

Sampath Tilakumara (sampath@lseg.com) is head of technology at LSEG Technology, Colombo, Sri Lanka.

Thayaparan Sripavan (thaya@lseg.com) is head of hardware-accelerated systems at LSEG Technology, Colombo, Sri Lanka.

Rasika Withanawasam (rasikaw@lseg.com) is senior software architect at LSEG Technology, Colombo, Sri Lanka.

AI Applications | DOI:10.1145/3343447

Technology Interventions for Road Safety and Beyond

BY C.V. JAWAHAR AND VENKATA N. PADMANABHAN

TOP PHOTO BY SNEHIT/SHUTTERSTOCK.COM

What hits a visitor to India first, quite literally, is the traffic. The combination of inadequate road infrastructure, increasing vehicle population, and poor driver training and discipline makes for a chaotic and often deadly mix. The result is a high rate of road accidents, with the estimate of fatalities ranging from one every four minutes[a] to over 238,000 each year.[b]

There is much ongoing work in academia, industry, and startups on using artificial intelligence (AI) and Internet of Things (IoT) technologies to improve the situation. The general goal is to have affordable technologies that work with humans through effective monitoring and feedback, rather than replacing humans through full autonomy.

The road and traffic conditions in India, which are quite different from those in the developed world, present interesting challenges (see Figure 1).

Road infrastructure. The road infrastructure has largely grown organically, without the benefit of long-term planning. It is of uneven quality, with safety hazards such as potholes, poor lighting, and inadequate signals and signage. Therefore, monitoring of the infrastructure to identify such hazards is quite important.

Vehicles. For cost reasons, vehicles often lack advanced features such as Advanced Driver Assistance Systems (ADAS). Also, the mix of vehicles tends to be very heterogeneous, spanning two-wheelers (for example, scooters and motorcycles), three-wheelers (for example, auto rickshaws), four-wheelers (cars), and larger vehicles (trucks, buses); in fact, it is not uncommon for even pedestrians to share the road space with vehicles. The heterogeneity in vehicle sizes and speeds often leads to a chaotic flow of traffic, far removed from adherence to lane discipline.

Drivers. Driving discipline is generally lacking, with drivers often cutting corners to "get ahead." Part of the reason for this is that

[a] http://bit.ly/31y18zX

[b] https://en.wikipedia.org/wiki/List_of_countries_by_traffic-related_death_rate

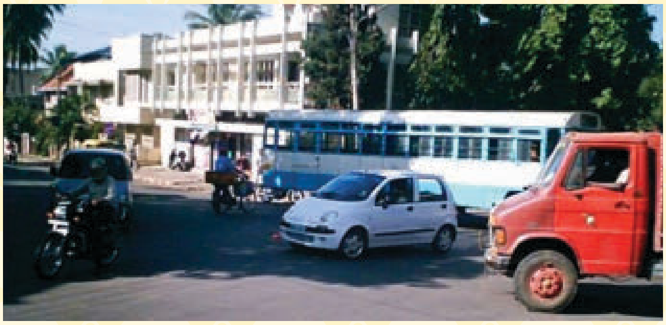


Figure 1. The traffic scene at a chaotic intersection with heterogeneous vehicles. Source: Nericell project; http://bit.ly/2MH6S6p.

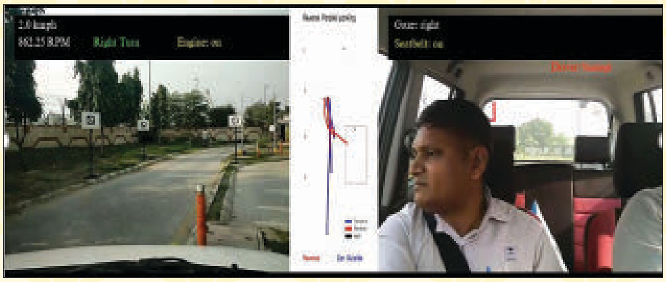


Figure 2. Tracing the trajectory during a parallel parking test (blue = forward, red = reverse) using a windshield-mounted smartphone, which analyzes outside and inside views concurrently. Source: HAMS project; http://bit.ly/2ZseQXC.

The general goal is to have affordable technologies that work *with* humans through effective monitoring and feedback, rather than replacing humans through full autonomy.

driver training and license testing lack thoroughness; by one estimate, 59% of driver licenses in India were issued without any test being taken at all.[c]

Research Examples

We touch on just a few examples of research inspired by this unique mix of conditions and constraints.

Datasets of Indian roads. With data being the fuel for AI research, there are efforts under way to assemble and release an Indian roads dataset.[d] This is helping to benchmark computer vision techniques on the unstructured Indian road conditions. It is also helping spur the development of new techniques for such data collection, such as low-cost inspection of road infrastructure (potholes, signage, and street lights) using computer vision and inertial sensing.

Driver training, testing, and assistance. There is much interest in using smartphones instead of special-purpose devices to monitor drivers and their driving, with a view to improving safety through effective feedback. A specific example is automated driver license testing, with tracking using a windshield-mounted smartphone in place of expensive infrastructure (like pole-mounted cameras) to make testing comprehensive and cost-effective (Figure 2).

Autonomous driving. While autonomous driving in India is likely far off because of the challenging road conditions, work is being done to enable autonomy for specific purposes in confined environments; for example, cargo vehicles at an airport.[e] There are also initiatives in place to spur research in autonomous driving more broadly through competitive grand challenges.[f]

Promising directions for ongoing and future work include technologies to aid traffic enforcement (for example, ticketing using automatic license plate reading), for pedestrians and two-wheelers (which dominate traffic and account for a disproportionate share of fatalities), and for effective simulation and what-if analyses.[g]

C.V. Jawahar is a professor at the International Institute of Information Technology Hyderabad, India.

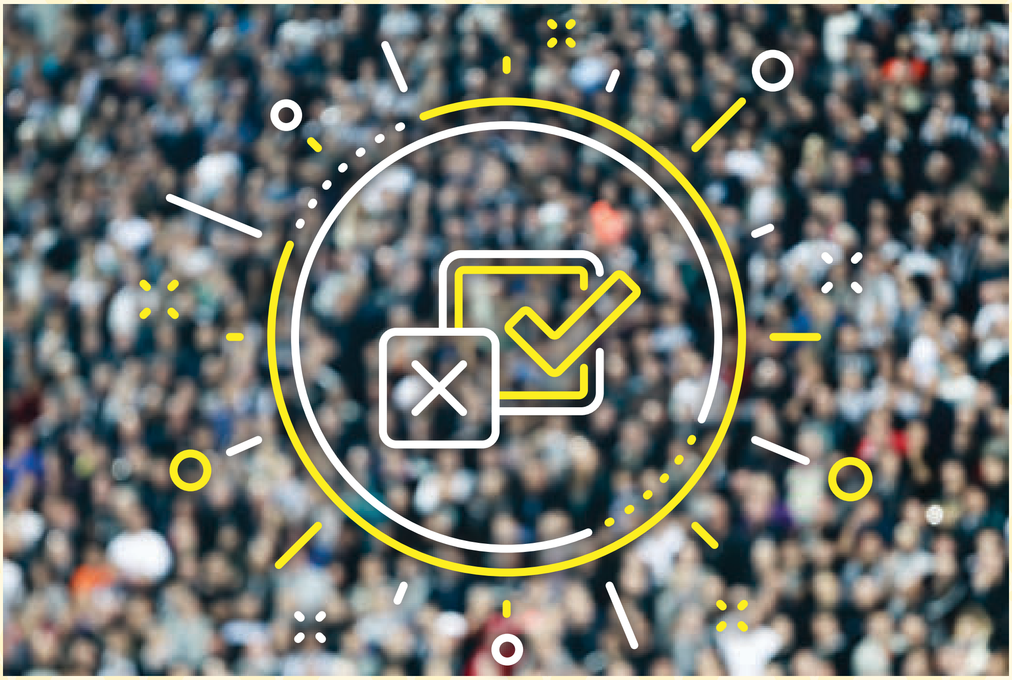
Venkata N. Padmanabhan is Deputy Managing Director at Microsoft Research India.

[c] http://bit.ly/31v5mZ4
[d] http://idd.insaan.iiit.ac.in
[e] https://www.atimotors.com/
[f] http://www.sparktherise.com/
[g] https://www.civil.iitb.ac.in/tvm/SiMTraM_Web/html/

JOB PLACEMENT | DOI:10.1145/3355268

Skill Evaluation

BY SHASHANK SRIKANT, ROHIT TAKHAR, VISHAL VENUGOPAL, AND VARUN AGGARWAL



UPWARD OF FOUR million graduates enter the labor market every year in India alone. India boasts of a large services economy, wherein a single company hires thousands of new employees every year. Meanwhile, product companies and small and medium enterprises (SMEs) look for a few skilled people each. This requires cost-effective and scalable methods of hiring. Interviewing every applicant is not a feasible solution.

On the other hand, graduates from 30,000+ institutes of higher education spread across 20+ Indian states face a constant challenge in signaling their competence to potential employers. Companies, most of which are located in the top 20 biggest cities in the country, bias their search by relying on proxies like university name and the city a college is located in. Applying such crude filters results in meritorious students from various demographics being ignored. Further, these students have no mechanism to get feedback on how their skills compare to those required by the industry.

Having systems that can intelligently and scalably assess a wide variety of skills is essential to addressing this broader problem affecting every modern-day labor market. *Aspiring Minds* was formed 10 years ago to address this challenging problem. We have developed a scalable platform to deliver standardized assessments to test job skills. The platform tests more than two million students every year and is used by 5,000+ companies including 100+ Fortune 500 companies.

A particular challenge in designing scalable assessment technologies is evaluating subjective, open-ended responses to questions.

A particular challenge in designing scalable assessment technologies is evaluating subjective, open-ended responses to questions. Such questions directly simulate a skill or a job task within the constraints of a test-like environment and are generally more informative than multiple-choice questions (MCQs). For instance, it is almost necessary to evaluate programming or spoken skills using such formats over MCQs. Evaluating such responses is an expensive, time-consuming process involving human graders, and suffers from standardization concerns. Automated grading has the potential to address these issues and impact millions of job seekers, trainers, and corporations.

At Aspiring Minds, we have, over the last decade, distilled a framework to cast the question of subjective assessments as problems in computer science, and specifically, in machine learning (ML).[1] In it, candidate responses are data points in a high dimensional space, from which we predict their true, latent, underlying score. This is a different paradigm altogether that we envision. While there

IMAGE FROM SHUTTERSTOCK.COM

exist solutions and products that evaluate language skills subjectively, most solutions provided by established, international educational testing and assessment organizations focus on testing general aptitude skills and adopt traditional testing formats like MCQs.

We illustrate the broad industry verticals we have developed tools for, each highlighting a research problem it addresses and the associated innovative intervention we devised.

▸ **Programming and software engineering.** Automata[6,8,9] uses ML models to automatically score computer programs on parameters such as functional correctness, complexity, and style. These models use intelligent features extracted from programs, which can signal correctness even when they fail to compile. Importantly, we designed them to be independent of the task the program solves, thus allowing to scale assessments to a wide variety of questions. There have been attempts by other research groups[2,7] at analyzing programs solving introductory programming problems. They, however, focus on providing automated feedback. Our work differs in that we focus on grading programs on a rubric. To achieve this, we extract key data flow properties in programs that capture their *meaning* and use them as features in an ML model; the problems we model are significantly more involved than introductory problems and exist in multiple languages.

▸ **Customer service.** The IT-enabled services (ITeS) market in India employs four million people and is a US$181-billion industry. Spoken English skills are central to this industry. *SVAR*[3,4] evaluates speaking skills at scale. Applicants call a phone number, have a conversation with an automated interactive system, and on hanging up, receive a score on their spoken skills such as pronunciation and fluency. It draws from speech and signal processing technologies and uses ML to predict these scores. To reduce evaluation time, and to improve model accuracy, we innovated by crowdsourcing parts of our feature extraction and model evaluation.

▸ **Blue-collar jobs.** Four-and-a-half million employees in India are estimated to be employed in blue-collar jobs. However, no automated means existed to assess motor skills, a key requirement in these jobs. Akin to how computers serve as a medium to test cognitive skills, we showed how touch devices can be used to assess motor skills.[5] This requires a person to use their fingers and wrists to play specific games designed for tablet apps. We have shown their performance on these tasks to correlate with on-job performance.

▸ **Professional communication.** Email correspondence has become an integral part of the communication tool chain in any organization. To test professionals' email writing skills, we employ deep learning and NLP to assess various aspects like grammar, content, and structure.

To our knowledge, this is the first attempt at designing and productizing such ML-driven technologies to assess these specific skills.

▸ **Domain knowledge.** In consultation with subject-matter and industry experts, we have designed 300+ tests for domain knowledge across various industry verticals such as IT, ITeS, retail, manufacturing, BFSI, hospitality, and telecom. Backed by statistical techniques such as item response theory, these tests provide standardized assessments in specific topics, helping create a level playing field for job applicants.

To our knowledge, this is the first attempt at designing and productizing such ML-driven technologies to assess these specific skills.

Over the years, we have gathered a database of applicants' performance in the various verticals discussed here. This has helped us quantify the state of employability in India, and study a year-on-year change in employability conditions. Since 2010, Aspiring Minds has released annual National Employability Reports, which have now become the gold standard for tracking the quality of higher education in India, aiding and informing policy formulation.

Besides these opportunities, we have also identified a number of challenges in using CS/ML for grading. These include issues around quality of labels (expert grades), low sample sizes, sample characteristics, standards for acceptable errors in models, among others. Several key issues are in developing models that are causal and addressing issues of fairness and bias in grading. These form areas of active research.

References

1. Aggarwal, V., Srikant, S., and Shashidhar, V. Principles for using machine learning in the assessment of open response items: Programming assessment as a case study. In *Proceedings of the Workshop on Data Driven Education*, 2013.
2. Gupta R.R. et al. DeepFix: Fixing common C language errors by deep learning. In *AAAI 2017*.
3. Shashidhar, V., Pandey, N., and Aggarwal, V. Spoken English grading: Machine learning with crowd intelligence. In *Proceedings of the 21st ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, KDD '15.
4. Shashidhar, V., Pandey, N., and Aggarwal, V. Automatic spontaneous speech grading: A novel feature derivation technique using the crowd. In *Proceedings of the 53rd Annual Meeting of the Association of Computational Linguistics and the 7th Intern. Joint Conf. Natural Language Processing*.
5. Singh, B.P. and Aggarwal, V. Apps to measure motor skills of vocational workers. In *Proceedings of the 2016 ACM Intern. Joint Conf. Pervasive and Ubiquitous Computing*.
6. Singh, G., Srikant, S., and Aggarwal, V. Question independent grading using machine learning: The case of computer program grading. In *Proceedings of the 22nd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*, 2016.
7. Singh, R., Gulwani, S., and Solar-Lezama, A. Automated feedback generation for introductory programming assignments. In *Proceedings of the 34th ACM SIGPLAN Conf. Programming Language Design and Implementation*, 2013.
8. Srikant, S. and Aggarwal, V. A system to grade computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining*, 2014
9. Takhar, R. and Aggarwal, V. Grading uncompilable programs. In *Proceedings of the Innovative Applications of Artificial Intelligence Conf. Assoc. Advancement of Artificial Intelligence*, 2019.

Shashank Srikant is a Ph.D. candidate at Massachusetts Institute of Technology, Cambridge, MA, USA.

Rohit Takhar is a research engineer at Aspiring Minds, Gurugram, India.

Vishal Venugopal is a senior software engineer at Aspiring Minds, Gurugram, India.

Varun Aggarwal is co-founder and Chief Technology Officer of Aspiring Minds, Gurugram, India.

R&D ROOTS | DOI:10.1145/3343443

Computing Research at Tata Consultancy Services

BY GAUTAM SHROFF AND K. ANANTH KRISHNAN

SOMETIME IN THE early 1960s a young general manager of the Tata Electric Co. in Mumbai (then Bombay) visited the nearby Tata Institute of Fundamental Research (TIFR),[a] where India's first electronic-stored program computer resided. That manager, F.C. Kohli, began using the computer to optimize load dispatch operations for the city. Within a few years J.R.D. Tata, then chairman and doyen of the Tata group of companies, called upon Kohli to look after the new computing division—Tata Consultancy Services (TCS). That division soon began catering to clients both outside the group and outside India—in Europe and North America—pioneering the 'offshore development' model.

The seeds of the Indian software industry had been sown.

In 2019, TCS revenues crossed $20 billion, ranking third after IBM and Accenture, and playing the role of a growth and transformation partner to large enterprises worldwide.[8]

In this article we focus on the unique role that research and innovation has played in TCS' journey from being the Tata group's computing division to its current place in global technology consulting. We shall highlight both some challenges faced along the way as well as lessons learned; the accompanying figure is a snapshot of this journey.

In 1981, TCS established a dedicated corporate research facility in Pune, the Tata Research, Development, and Design Centre (TRDDC).[b] It was headed by E.C. Subbarao, a prominent materials scientist from IIT Kanpur, who began applying computational materials engineering for Tata Steel, which then dominated the group's business. Indeed, research in TCS was joined at the hip with business from its inception.

Soon Kesav V. Nori joined TRDDC from TIFR and began adapting compiler technology to TCS' fledgling software services business, successfully automating many of the conversion projects that TCS won through the 1980s, for example, from one language/database to a more modern platform.

The seeds of the Indian software industry had been sown.

PHOTO BY BILAL ALIYAR M/SHUTTERSTOCK.COM

In the late 1980s, TCS was awarded a large, complex development project by SEGA, the Swiss financial depository, clearing and settlement organization. Such a project demanded computer-aided software engineering (CASE)[3] tools, with research supplying the technology to capture system-level specifications from which code could be generated automatically. Later, in the early 1990s, Kohli felt there was an opportunity to develop a product for the Swiss private banking industry, and charged research with creating a new generation of CASE tools—MasterCraft.[12] By employing code-generation, the banking product remained insulated from multiple generations of technology change. Today, .NET, Java, and Web versions of the product (TCS BαNCS[9]) have also been instantiated even while its banking functionality evolved independently.

With the turn of the century, TCS research expanded beyond its core location Pune as well as its traditional areas: Security and bioinformatics groups were seeded in Hyderabad,

[a] Tata Institute of Fundamental Research is a research laboratory funded and run by the Government of India.

[b] TRDDC is now part of TCS Research.

embedded systems in Bangalore and later Kolkata, and architecture in Delhi. These new groups quickly began impacting TCS business: For example, when TCS launched iON,[11] a hosted application platform for the SMB segment in India, a runtime configurable multi-tenant architecture[6] developed by the architecture research group served as its initial technology base.

The 2000s saw a rapid growth in infrastructure management services, where TCS would manage entire datacenters and networks for large customers. This was traditionally a people-intensive exercise, but also turned out to be an excellent target for researchers to apply a variety of AI/ML techniques for automation. After many years of field deployment 'under the radar,' IGNIO,[10] an AI-driven enterprise automation product developed in TCS Research, was formally launched in 2015.

Today, TCS Research covers a wide variety of areas (as listed in the figure) and is increasingly multidisciplinary. Many long-term investments are now actively applied to TCS' business; for example, marrying materials science research with machine learning to develop digital twins to optimize industrial operations,[5] and with knowledge ontologies to complete the manufacturing 'digital thread.'[2] Or combining metadata abstractions from our earlier software engineering tools with deep reinforcement learning to optimize supply chains and drive personalization in customer interactions.[1] Further, TCS eats its own dog food—it has deployed a home-grown enterprise social media platform[7] and more recently a deep-learning-based conversational system across all its 400K+ employees.[c]

There have been many lessons learned along this journey; we mention some critical ones here: First, research initiatives have always preceded their applicability, and so continuing to invest in research areas seemingly unrelated to the current business pays off in initially unforeseen ways. For example, research in genome-based early prediction of rare diseases[4] later enabled TCS' business to build genome analysis pipelines for pharma customers. Deep expertise in computational chemistry[3] is now allowing TCS to design new chemical formulations and molecules for customers, a very different kind of service that could potentially expand the very scope of its core business in the future.

Second, translating research results into business outcomes not only requires careful shepherding, but also for research to evolve along with the business of the company: Just as IBM Research had to develop 'services science'[d] to support its emerging global services business, so too has TCS Research evolved, from initially using computer science to accelerate software development and IT tasks, to now applying a combination of both AI and domain research to transform product engineering, operations, and business models for its customers.

Last but not least, nurturing a vibrant research environment within a large and often independently successful core business becomes a challenge in itself. For many years, research was viewed as a luxury indulged in because one could afford it, that is, the foresight of investing in research was far from widely shared. Only hindsight shows indulgence paying off as the world itself is changing so rapidly, with every business increasingly becoming a technology business, across industry verticals.

To conclude, we submit the role of research for players in the technology services industry, such as TCS, is to act as the bridge between fundamental scientific advances (in computing and beyond) and transformative business ideas and product innovation for large enterprises that form their customer base.

TCS Research Areas

- Physical Sciences
- Software Systems and Services
- Life Sciences
- Embedded Systems and Robotics
- Cybersecurity and Privacy
- Computing Systems
- Behavioral, Social, and Business Sciences
- Data and Decision Sciences
- Deep Learning and AI
- Computing Foundations
- Media and Advertising

Timeline

1981
Establishment of TRDD, Pune; research focus on computational engineering to support Tata group.

1983 → 1990s
Automation of migration to code generation tools; creation of MasterCraft.™

2000s
Expansion of research beyond Pune to several new centers and focus areas with TCS business expansion beyond software development into engineering and infrastructure services, cloud computing, and process outsourcing …

2008
The growing importance of data, emergence of big data, maturing of AI and deep learning for enterprise applications …

2015
Launch of Ignio™

2018
Creation of a theory group, focus on ad-tech, media, and so on.

… the evolution continues …

Evolution of TCS research.

[c] https://on.wsj.com/2R4xflP

[d] http://bit.ly/31oAXM0

References

1. Barat, S. et al. Actor-based simulation for closed loop control of supply chain using reinforcement learning. In *Proceedings of the 18th Intern. Conf. Autonomous Agents and MultiAgent Systems. Intern. Foundation for Autonomous Agents and Multiagent Systems*, 2019, 1802–1804.
2. Gautham, B.P., Reddy, S., Das, P. and Malhotra, C. Facilitating ICME through platformization. In *Proceedings of the 4th World Congress on Integrated Computational Materials Engineering*. Springer, Cham, 2017, 93–102.
3. Gupta, R., Dwadasi, B.S., Rai, B. and Mitragotri, S. Effect of chemical permeation enhancers on skin permeability: In silico screening using molecular dynamics simulations. *Scientific Reports 9*, 1 (2019), 1456.
4. Punwani, D. et al. Multisystem anomalies in severe combined immunodeficiency with mutant BCL11B. New England *J. Medicine 375*, 22 (2016), 2165–2176.
5. Runkana, Venkataramana. Model-based optimization of industrial gas-solid reactors. *KONA Powder and Particle J. 32* (2015): 115–130.
6. Shroff, G., Agarwal, P. and Devanbu, P. Instant multi-tier Web applications without tears. In *Proceedings of the 2nd India Software Engineering Conf.* ACM, 2009, 3–12.
7. Singh, M. et al. KNADIA: Enterprise KNowledge assisted DIAlogue systems using deep learning. In *Proceedings of the IEEE 34th Intern. Conf. Data Engineering*. IEEE, 2018, 1423-1434.
8. TCS Annual Report 2018–2019; https://on.tcs.com/2Ivq8Py
9. TCS BαNCS; https://www.tcs.com/bancs
10. TCS Ignio; https://www.digitate.com
11. TCS iON; https://www.tcsion.com
12. TCS MasterCraft; https://mastercraft.tcs.com

Gautam Shroff (gautam.shroff@tcs.com) heads TCS Research and is based in Delhi, India.

K. Ananth Krishnan (ananth.krishnan@tcs.com) is Chief Technology Officer for TCS and is based in Chennai, India.

DOI:10.1145/3347863

BY PANKAJ JALOTE AND PARI NATARAJAN

The Growth and Evolution of India's Software Industry

THE DEVELOPMENT OF the Indian software industry is an archetype of how economic liberalization combined with an entrepreneurial spirit can build an industry that today contributes as much as 8% to the GDP of a fast-growing country like India. On the back of thousands of IT services companies that were built over the last three decades, the industry has generated US$177 billion in revenue and more than US$135 billion in exports in FY 2018–2019 alone. The IT industry has also created over four million direct jobs and 12 million indirect jobs in India. A testament to this growth is the fact that the largest Indian IT services company is currently valued at over US$100 billion and generates over US$20 billion in revenue.

Over the years, the Indian software industry has matured from providing cost-effective back office support to driving the digital transformation agenda ahead in global companies. Increasingly, leaders of more than a thousand global enterprises across the U.S., Europe, and other locations have realized India's potential and have set up their own IT or R&D centers to take advantage of the vibrant Indian software ecosystem.

The current wave of Indian software entrepreneurs is focusing on building platforms and products for Indian and global markets. This has led to the creation of more than 7,000 tech start-ups in India. India is already home to 18 unicorns (start-ups valued in excess of US$1 billion), and another 10 are expected to be added by the end of 2020.

The Indian software industry has accelerated the adoption of digital technologies in the country. The industry has played a crucial role in providing digital identities to over one billion people in the country, which is further enabling the provision of services across industries such as banking, healthcare, and education in an efficient manner. The next generation of Indian software companies is helping millions of small and medium businesses (SMBs) and individual workers such as cab drivers and delivery personnel move into the formal economy.

This article is not just a story of the Indian software industry but also of the entrepreneurial capability of India's vast talent pool.

Growth of the Software Industry in India

For the purpose of discussion, the growth and evolution of the industry can be viewed in three broad phases:

▸ **Pre-2000 era: The growth of software exporting firms.**

For the two decades in this period, the software sector was largely comprised of firms looking to provide software services to global clients. The focus was on exports, and most companies viewed themselves as software exporters. The companies started solving Y2K issues for their customers and further extended their offerings to

help companies manage their legacy portfolio of applications and infrastructure. The first wave of the global Internet and dot-com era created intercontinental Internet infrastructure. Indian companies were able to leverage this infrastructure to deliver software development-related services to global enterprises remotely.

Realizing the potential and the availability of talent, some multinational corporations established their own offshore development centers in India. Companies involved in the software aspects of hardware—for example, design of tools or VLSI (very large-scale integration)/system design—also took root, diversifying their services portfolio.

▸ **Circa 2000–2010: The rise of Indian software multinationals and R&D centers.**

With experience in dealing with complex IT systems and confidence in working with international customers, several companies became multinationals with offices and centers across countries. They offered a wider range of services like executing large and complex projects involving integration, complete end-to-end solutions including management of IT infrastructure, running the services, providing IT strategy, and other related services.

Global multinational companies also realized India's potential in software services and started increasing their direct presence in India by setting up IT, business process management (BPM), and R&D centers. To date, 1,250 companies from around the world have set up their own centers in India across almost all key industry verticals. Software/Internet, telecom, semiconductor, automotive, and industrial are the top industries present, with R&D being a strong focal

IMAGE FROM SHUTTERSTOCK.COM

Figure 1. Growth of IT services, GICs, and tech start-ups in India.

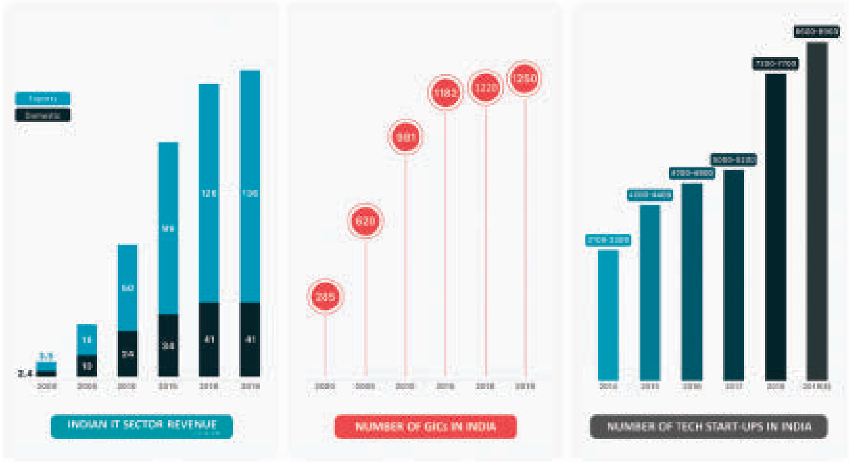
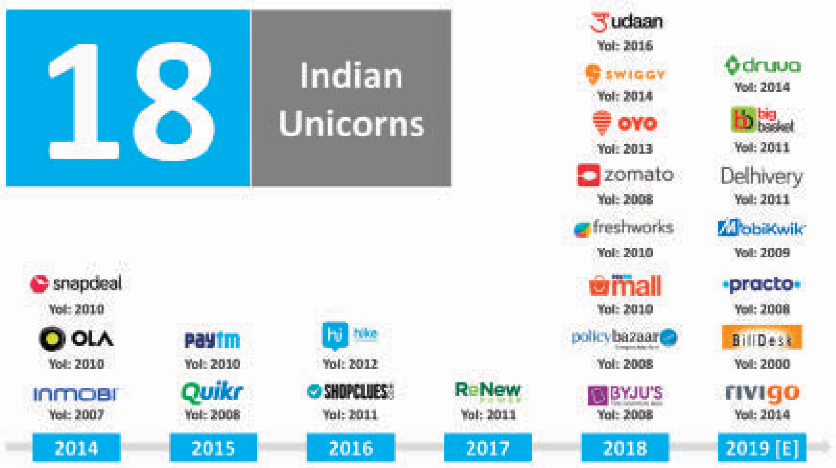


Figure 2. The rise of unicorns in India.



point. Enterprises across industries such as banking, retail, and healthcare also started driving digital engineering work from their India development centers.

Today, several centers have matured to deliver end-to-end products from India. These centers also act as the gateway to Asia, helping with product localization and creation of new products for these markets. Even next-generation companies have started setting up centers in India. Uber set up an engineering center in 2017, and OVH—a unicorn from France that provides cloud services—set up an R&D center in the country last year.

Over 400,000 engineers work in global R&D centers in India. Bangalore, Pune, Hyderabad, National Capital Region (Delhi, Noida, Gurgaon), and Chennai are key locations for such centers, amplifying the possibility of ecosystemwide learning, relearning, innovation, and partnership.

▸ **Circa 2011 to present: Vibrant and innovation-driven multi-dimensional sector.** The Indian software ecosystem has now evolved into an extremely dynamic and varied sector that is building and managing the most complex IT systems for global enterprises. The combination of available talent, lower rates of brain drain to the U.S., the presence of large technology companies' R&D centers, and the presence of global venture capitalists has helped accelerate the growth of the start-up ecosystem. India, today, has over 7,000 start-ups (started less than five years ago), and over 1,200 technology start-ups were established in just the last year.

There are largely two types of technology start-ups. The first are consumer-led and largely focused on the India market. Initially these were replicas of U.S. companies, but soon morphed with unique innovations for the India market. For example, the cash on delivery model in e-commerce was pioneered in India and is now used globally. The second set of start-ups are focused on serving the U.S. and European markets.

In the last few years, 18 start-ups touched US$1 billion in market capitalization. Walmart bought India's largest e-commerce company, Flipkart, which is only about 11 years old, at a valuation of US$21 billion. OYO Rooms, a technology-enabled franchise model hotel chain, was started by a 20-year-old, and now has the largest number of rooms under management in India, overtaking both traditional Indian and global hotel chains.

Start-ups are driving innovation at an accelerated pace. To maintain the warp speed of innovation, large companies are building partnerships with the start-ups and are actively looking at acquisitions, both for talent and intellectual property.

Impact of IT Industry on India

The IT industry's impact on India is profound. It is a positive contributor to India's revenue growth, talent capability, diversity in workforce, and its digital infrastructure.

Growth in exports. In terms of revenue and foreign exchange, this sector has transformed India's finances, and is effectively financing a large share of imports. The sector is currently the largest forex earner from exports and accounts for over 25% of the country's total exports. The sector is already contributing over 7.9% to India's GDP.

Capability development and employment creation. No other industry segment has generated as many jobs for the middle class. The sector directly employs over four million people and indirectly supports an additional 12 million jobs. The industry was also a major trigger for the government to push for an increase in output of engineering colleges to over 700,000 graduates a year.

Companies have also set up processes to hire, train, and engage thousands of employees. In fact, Indian IT services companies spend over US$1.6 billion a year on employee training. Large technology companies have set up campuses exclusively focused on training their employees on skills relevant to their global customers. Over 500,000 engineers in India are already equipped with relevant digital skills to drive digital transformation. FutureSkills, an initiative of the National Association of Software and Services Companies (NASSCOM), has an ambitious goal of training another two million people in digital technologies over the next few years.

The extensive engineering education system and the deployed talent pool in the IT industry are also helping improve the digital capabilities of Indian enterprises. The technical and managerial talent from IT companies have moved to Indian enterprises to help them accelerate their digital transformation initiatives.

Female empowerment. The industry has been supportive of women in the workforce, an aspect where India has traditionally lagged. Some 30% of the IT sector workforce is comprised of women employees and this has been a trend since the early stages of its development. The sector has not only helped empower women but has also provided them with highly aspirational career options.

Start-up ecosystem. The start-up ecosystem in India attracted over US$10 billion in investments from venture capitalists from across the world between 2016 and 2018. US$6 billion has already been invested in Indian start-ups by SoftBank out of its US$100 billion Vision Fund.

Start-ups such as Flipkart, Ola, and Swiggy have helped create or digitally enable millions of jobs such as cab drivers and e-commerce/food delivery professionals. These companies are also empowering the country's 60 million small and medium businesses by digitally enabling their operations. Start-ups such as Power2SME and CapitalFloat are offering innovative financial services for SMBs, including "flow-based lending;" a lending model that provides credit to SMBs based on an analysis of their financial transactions, thereby improving SMBs' ability to invest and grow their businesses.

Digital infrastructure. Within a span of about a decade, Indian IT companies have taken several services being provided to citizens and corporations and moved them online. Most of these systems have been developed by indigenous IT companies, and many are also maintained and managed by them. Examples include the Ministry of Corporate Affairs system for corporate tax filing, the income tax management system, including e-filing of tax returns, the entire India Stack digital infrastructure, the Goods and Services Tax system, the passport system, the Indian rail reservation system (that books over 200 million tickets annually), the Aadhaar unique identification infrastructure—the largest in the world (whose chief conceptualizer and first CEO, Nandan Nilekani, is a product of the IT industry), and others.

India's global perception. Finally, it should be noted the software sector has perhaps played the most crucial role in changing the global perception of India. Until the 1980s, India was perceived as a poor country that needed support from more developed nations. Today, this view has changed, and India now has a seat at the global table. The world is aware of India's technology prowess and is actively looking to make investments, form partnerships, and tap India's bustling technology ecosystem. Frequent foreign travelers can attest to the fact that the quality of interaction with local people has evolved dramatically over the last quarter century due to the IT industry's widespread impact.

Figure 3. Local impact of Indian software sector.

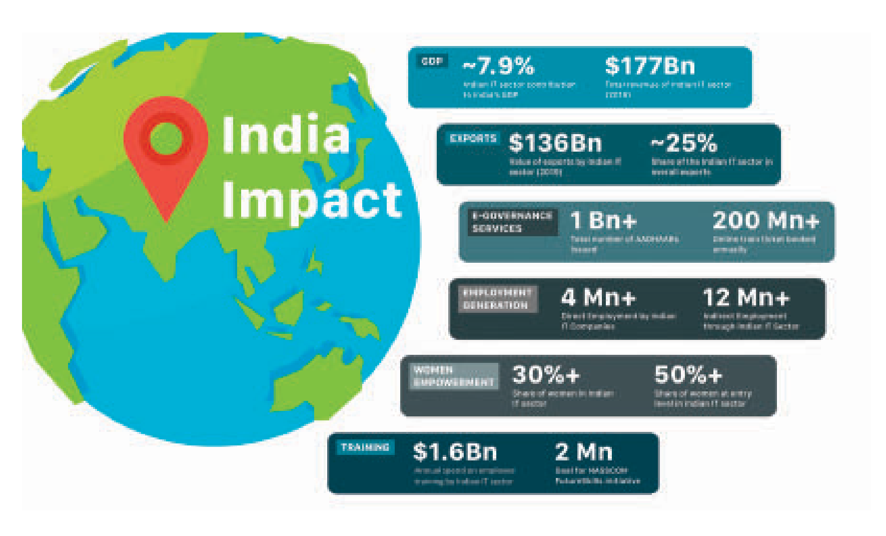
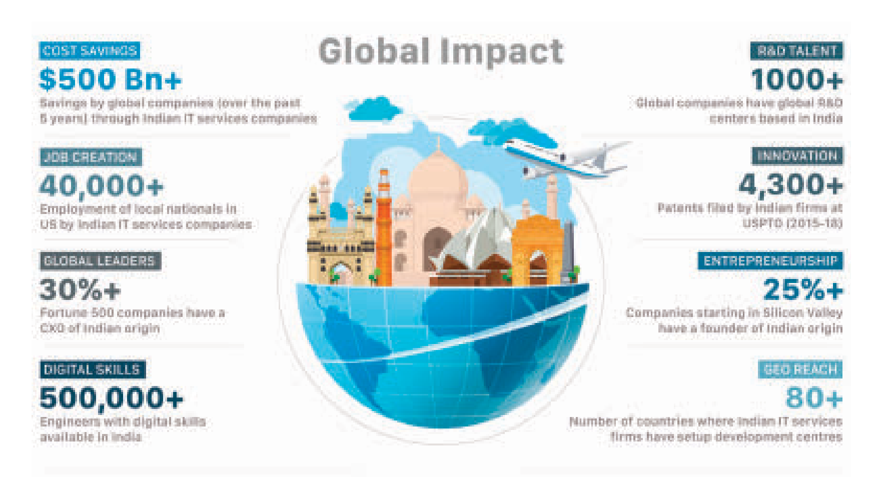


Figure 4. Global impact of Indian software sector.



The Indian IT sector is in a unique position to lead global technology innovations over the next decade.

Global Impact

Rising from a position where it was almost impossible for a poor and impoverished country like India to create capabilities around software technology, India has emerged as a software powerhouse serving the world.

The IT sector has helped global corporations optimize their cost, improve quality, create jobs and capabilities worldwide, and drive global business models and technology innovations.

Cost optimization and quality. India's high quality and relatively lower cost engineering, IT, and BPM talent has helped corporations gain huge cost savings that have allowed them to increase their shareholder returns and invest in growth and transformational initiatives. Assuming a cost difference of over 50% between developed locations and India, the Indian IT industry has helped global enterprises save over US$500 billion in the last five years alone.

It was important for the Indian IT companies to ensure that they could deliver high quality software. In fact, they were among the first to adopt the software development process standard called Capability Maturity Model (CMM), launched by the Software Engineering Institute at Carnegie Mellon University in 1987. By 1998, half of all global companies assessed at CMM level 4 or 5 were from India. Companies from across the world, and even countries wanting to develop their own ITsectors, turned to India's experience of rigorous software development process and the use of quantitative techniques to manage them. Indian companies shared what they learned at software-related conferences, workshops, seminars, and so forth. Delegations from various countries visited India to learn from its experiences, and case studies were also developed. There were books like *CMM in Practice* (Addison-Wesley, 2000), and *Software Project Management in Practice*, (Addison-Wesley, 2002), which were translated into other languages including Chinese, Japanese, Korean, and French.

Global delivery model. Riding on its CMM capabilities, Indian IT companies have pioneered the global delivery model where distributed teams can seamlessly work together to deliver complex software projects.

The global delivery model innovation has quickly become an industry best practice among all global IT services companies as well as enterprises. Large enterprises have anchored on the global delivery model to set up their own IT and R&D centers in India to drive IT and product innovation. Today, over 500 global companies have their second largest development centers based in India. Not just American and European countries, but also companies from Asian locations including China, Indonesia, South Korea, and Japan have established a presence in India. Companies such as IBM and Accenture have increased their headcounts to around 100,000 in India, due to their ability to use a global delivery model to execute complex projects for customers.

Global innovation and digital transformation. Large global companies have realized the capability of Indian talent and started focusing on driving core innovation from the country. In fact, several companies have started filling global patents for ideas that were conceptualized and productized from their India centers. Between 2015 and 2018, more than 4,300 patents were filed by India-based companies/offices in the U.S. Patent and Trademark Office. Companies have also expanded the roles at their India centers to include product management, customer success, sales, and marketing. Some companies, including Cisco and Samsung, consider their India centers as second headquarters. Also, India is rapidly rising as a hub of global Centers of Excellence (COEs) for modern technologies such as artificial intelligence/machine learning, Internet of Things, robotic process automation, and others.

The Indian software industry was quick to realize the changing needs of customers and started supporting global digital transformation initiatives of several legacy companies in the retail, manufacturing, energy, and utilities industries.

The large Indian IT companies are now multinationals in their own right—the top companies operate in over 50 countries, where they have substantial offices. Many have significant development centers in other regions like the United States, Mexico, China, Europe, and Australia, employing thousands of software engineers and managers. The Indian IT services industry is estimated to employ a total of 40,000+ locals in the United States alone. Additionally, the industry is also exporting its massive talent training infrastructure to global locations. Tata Consultancy Services has set up a training hub in Cincinnati and is creating a pipeline of graduates coming out of U.S. universities.

The worldwide impact of the Indian software industry is widely evident:

- Code written by Indians is present in almost all systems with software, including cars, consumer electronics, enterprise software solutions, industrial products, banking systems, and more;
- Indian designers are involved in most chip and system designs by major multinationals;
- Indian IT firms have development centers in over 80 countries around the world; and
- Over 1,000 companies develop global products from their centers in India.

Key Lessons

Observing the growth and impact of the Indian IT industry provides a set of valuable lessons that can be replicated for the development of other industries in India and other countries.

Government involvement. Minimal government interference coupled with supportive incentive policies was a key success factor. The Indian government did not regulate the industry and created tax incentives for both importing technology and for revenue from exports.

Skilling and development. Focus on skills and talent development has been instrumental in the growth of the IT industry. Even in their early stages, IT companies spent significant time and money developing the skills of their employees ahead of time. This helped companies rapidly address the changing technology needs of their global customers.

Process orientation. The heightened focus on process orientation in the Indian software industry has undoubtedly contributed to its meteoric growth and has also enhanced the perception of the industry globally. A strong process and continuous improvement focus is a catalyst for both quality and productivity.

Industry collaboration. One key factor in the success of the Indian IT sector has been its ability to bring companies together to develop an industry. The sector has created a huge collaboration ecosystem in the form of an industry body—NASSCOM. The association has helped develop best practices that get disseminated to companies across the sector.

Scale and entrepreneurship. Widespread industry effort to promote, cultivate, and celebrate entrepreneurship has created an ecosystem for entrepreneurs to conceptualize, fund, and scale IT companies. The first generation of entrepreneurs focused on building and scaling IT services companies, while the second is focusing on building IT products and IP-led services companies.

The Way Forward

The Indian IT sector is in a unique position to lead global technology innovations over the next decade.

The Indian education infrastructure is being rapidly overhauled by dedicated government initiatives. The government has announced the creation of 17 new Indian Institutes of Technology (IIT) across India to further improve the quality of engineering education. IITs have some of the country's best engineering faculty and education infrastructure. The millions of engineers who will graduate in the next few years will be adept at machine learning, cloud computing, and other new-age digital technologies. As a result India will continue to be a source of skilled digital talent and intellectual property for more than 2,000 global enterprises.

India's per capita income is expected to cross US$3,500 by 2025 from the current US$2,000. This will increase discretionary spending by the population, creating a huge consumer market, potentially triggering the next wave of digital entrepreneurs building India-focused technology. Venture capital activities will increase due to massive domestic opportunities and the ability of Indian start-ups to build global products. The availability of capital will catalyze the creation of many companies with valuation in excess of US$1 billion across India.

The government is expected to accelerate the creation of public digital infrastructure to streamline existing citizen services and create new services. The India Stack model will be expanded to create industry-specific initiatives in areas such as healthcare, supply chain, and education. This will result in technology getting weaved into the fabric of the Indian workforce across agriculture, healthcare, education, and other industries. Millions of digitally enabled jobs and job categories will be created in the process.

Further, second- and third-tier locations will join India's software ecosystem due to the strong mobility network, education, and digital infrastructure built over the last decade. Global companies, Indian IT companies, and start-ups will leverage these cities to drive innovation.

Over the last three decades, India has risen as a technology and software trailblazer, and with concerted efforts by the entire ecosystem including Indian IT companies, multinationals, start-ups, and the government, India has the potential to further establish its standing as a world leader in the software sector.

Pankaj Jalote (jalot@iiitd.ac.in) is Distinguished Professor (and founding director 2008–2018) at Indraprastha Institute of Information Technology (IIIT), New Delhi, India.

Pari Natarajan (pari@zinnov.com) is the Chief Operating Officer of Zinnov, Bangalore, India.

Suggested Reading

NASSCOM, *Strategic Review: IT-BPM Sector in India 2019: Decoding Digital*

NASSCOM, *Future Skills—A NASSCOM Initiative*

NASSCOM, *Women 'in'Equality—Not Anymore!*

NASSCOM-Zinnov, *GCC 3.0: Spotlight on Digital, Partnerships, New Delivery Models & Future Skills*, 2019.

NASSCOM-Zinnov, *Indian Tech Start-Up Ecosystem 2018: Approaching Escape Velocity*

Press Information Bureau, Government of India, Ministry of Electronics & I (May 3, 2017), *Employment Prospects in India's IT Sector: Robust Outlook*

big trends

DOI:10.1145/3343456

BY PUSHPAK BHATTACHARYYA, HEMA MURTHY, SURANGIKA RANATHUNGA, AND RANJIVA MUNASINGH

Indic Language Computing

IN APRIL 2019, following the Easter Sunday bomb attacks, the Government of Sri Lanka had to shut down Facebook and YouTube for nine days to stop the spreading of hate speech and false news, posted mainly in the local languages Sinhala and Tamil. This came about simply because these social media platforms did not have the capability to detect and warn about the provocative content.

India's Ministry of Human Resource Development (MHRD) wants lectures on Swayam[a] and NPTEL[b]—the online teaching platforms—to be translated into all Indian languages. Approximately 2.5 million students use the Swayam lectures on computer science alone. The lectures are in English, which students find difficult to understand. A large number of lectures are manually subtitled in English. Automatic speech recognition and machine translation into Indian languages will be great enablers for the marginalized sections of society.

Requirements like these are real and abundant. These are social and commercial needs, whose servicing requires user interaction and information dissemination in languages other than English. Only around 10% of India's population, or about 125 million people, can speak English; only about half that number is comfortable reading and writing in that language. The social media activity of the youth of the Indian subcontinent (where 65% of the population is below the age of 35) generates a huge amount of e-content, much of which is in text form, is multilingual, and even code-mixed (text in multiple languages at the same time, often in Roman script). The numbers are mind-boggling:[c]

- 462.1 million Internet users (34% of the population; the global average is 53%).
- 430.3 million users access the Internet via mobile devices (79% of total Web traffic).
- 250 million social media users (19% of the population; the global average is 42%).
- 260 million WhatsApp users, and 53 million Instagram users.

Sri Lanka alone has seven million Internet users (2018 data), which equates to a penetration of 32%.

There is no doubt that speech and natural language processing (NLP) of Indic languages is hugely important and relevant, and has the potential to influence the lives and activity of at least 20% of the world's population.

Challenges of Indian Language Computing

The Indian subcontinent is divided into seven independent countries: India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka, and the Maldives.

There are approximately 1,599 languages in India, out of which about 420–440 are in active use. Languages in the region fall into four major linguistic groups: Indo-Aryan (spoken mainly in the northern part of south Asia and in Sri Lanka), Dravidian (spoken mainly in south India), Tibeto-Burman (spoken mainly in northeast India), and

[a] https://swayam.gov.in/

[b] https://nptel.ac.in/

[c] *India Today*, April 2018 issue.

IMAGE BY JOAT

Diversity is the name of the game for Indic-language computing; shown here are scripts in Devanagari, Brahmi, Odia, Tamil, Telugu, Malayalam, and Sinhala, among other languages.

Austro-Asiatic (Khasi in Meghalaya, and Munda in Chhotonagpur). These language families each have their own linguistic characteristics, whose richness and complexity have been delved into in multiple scholarly treatises.[11] These complexities, along with techno-human constraints, give rise to the challenges of Indic language computing, some of which are described here.

Scale and diversity. For Indic languages, solutions must be simultaneously proposed for multiple languages. There are 22 major languages in India, written in 13 different scripts, with over 720 dialects. There is a need to develop approaches that are generic, and scaling to multiple languages should be only a task of adaptation. As the languages are quite different, there is a lot of effort required to arrive at common solutions. Although E2E (end-to-end) is the buzzword today, use of multiple scripts for Indian languages makes systems complex (as illustrated in the accompanying figure).

Long utterances. Indian-language utterances are much longer in duration compared to English, and hardly contain punctuation. A typical English sentence has about 70 characters, while a sentence in an Indian language typically averages 130 characters. E2E systems perform poorly with long sentences.

Code mixing. Code mixing is the use of more than one language in text/utterance. Handling code switching from one language to another in both automatic speech recognition (ASR) and text to speech (TTS) is a challenge. In ASR, the language boundary could be an important cue for semantics (assuming the lexicon accounts for the vocabulary of both languages). Also, Indian language words are included in an English sentence, where gerundification (such as "I'm chalaaoing a car," meaning "I am driving a car") of Indian-language nouns is common. In TTS, producing code-switched systems requires the prosodic characteristics of the language and the speaker are preserved, especially when code switching involves stress-timed and syllable-timed languages. The interplay between languages in terms of prosody needs to be understood to make the sentences sound natural.

Resource scarcity. Indic-language computing is bogged down by paucity of data. Language computing these days is primarily data-driven, with sophisticated machine learning techniques employed on the data. The success of these approaches depends crucially on the availability of large amounts of high-quality data. We take the example from automatic machine translation (MT), which is highly data-driven these days: the Hansard corpus for English-French contains 1.6 billion words; the Europarl Parallel Corpus for 21 European languages contains about 30 million words; WMT 15 data for English-Czeck contains about 16 million parallel sentences; and WMT 14 data for English-German contains about 4.5 million parallel sentences. An Indic-language example with comparable size is the CFILT-IITB English-Hindi corpus, which includes 800,000 parallel sentences.

Other languages offer very little language data. For example, available parallel corpora for Sinhala-Tamil are well below 50,000 sentences. Even raw, clean corpora are of great value for language computing. Modern-day deep learning techniques start with word embeddings (WEs). WEs are learned from huge amounts of corpora (millions of words) that capture the context distribution for words and phrases. Such distribution captures semantics, which is an elusive entity, computationally speaking. Many Indic languages do not have a processable clean corpus from word lists, WEs, and a rich lexicon can be built. Another application area that is affected by paucity of data is ASR-TTS. Spoken signals must be correct, with proper text units. Then there are transcriptions of spoken utterances that need to be accurate. Although there are subtitled YouTube videos and lectures, they require curation, as time alignments are quite poor. However, the number of available hours of training data is small, leading to poor alignments.

Absence of basic speech and NLP tools. The NLP pipeline starts with word-level processing, and goes all the way up to discourse computation (connecting many sentences together with attention to coherence and cohesion).[2] The tools used at each stage of this pipeline are affected by the accuracy of tools in the preceding stages. For English, since many groups across the world have worked on the computational processing of the language, a staged development of NLP tools of English occurred. NLTK,[d] a GATE-like[e] NLP framework came into being, paving the way for large application development in English. In contrast, even basic morphology analyzers that split words into their roots and suffixes do not exist for most Indic languages, and even if they exist, their accuracy level is low.

Absence of linguistics knowledge. Though speech processing and NLP are data-driven, linguistics insight and understanding of language phenomena often help solve the problem of accuracy saturation. Deep understanding of language phenomena helps design

[d] https://www.nltk.org/

[e] https://gate.ac.uk/

good problem-solving strategies, and helps immensely in error analysis and explainability. Many Indic languages do not have a linguistics tradition.

Script complexity and non-standard input mechanisms. In an Indic language such as Devanagari, there are 13 vowels, 33 consonants, 12 vowel marks or matras, complex conjunct characters, and special symbols such as anusvara, visarga, chandra bindu, and Nukta.[f] This makes input speed slow (8–10 words per minute, compared to 20–30 w.p.m. in English). Though an InScript keyboard layout has been mandated by the Government of India, there are questions on its optimality and ease of use. Suggestions for more efficient keyboard layouts keep appearing. The problem is compounded by the presence of 13 different scripts, which drives people to resort to Roman input through transliteration most of the time.

Non-standard transliteration. There are variations in representation when it comes to transliteration in Roman. For example, the Hindi word for "mango" (a fruit) can be transliterated as "am," "Am," or "aam." This creates a challenge for processing, and does not help the English-illiterate.

Non-standard storage. The appearance of Unicode for Indic languages and its adoption as the standard encoding of Indic language e-content was rather slow. As a result, many proprietary fonts exist, and the content of those fonts require downloading and algorithmic adaptation.

Man-made problems. Problems are further compounded by the fact that noise levels on the subcontinent average about 70dB, while the maximum permissible level is about 55dB. This challenges speech recognition technologies.

Some challenging language phenomena. A language phenomenon across major Indian languages is compound verbs (CVs), whose processing is a must for Indic-language NLP (INLP). CVs are composed of two verbs such that the main information content of actual action is carried by the first verb (called the polar) and the Gender-Number-Tense-Aspect-Modality (GNPTAM) information are marked on the second verb (called the vector). Elaborate machinery is needed for computational processing of CVs, starting from morphology, and up to the pragmatic level.[3] As an illustration, consider the Hindi compound verb:[g]

H_1: *bol uthaa* (Hindi string)

G_1: *speak rose* (gloss)

T_1: *spoke up* (English translation)

There is a sense of abruptness/urgency/letting-out-pent-up-feeling that is an additional layer of meaning carried by the vector verb on top of the main action of speaking (the polar). Catching such fine nuance is essential, for example, in sentiment and emotion analysis.[8]

Morpheme stacking. Many Indian languages show heavy stacking of morphemes (for the example, subscript 2 means the second sentence in the document):

M_2: *gharaasamorchyaanii malaa saaMgitle* (Marathi sentence).

P_2: *ghar+aa+samor+chyaa+nii+malaa+saMgit+le* (showing morphemes).

G_2: *house+<morpheme: oblique marker>+front+of+<ergataive marker: agent> me told* (gloss).

T_2: The one in front of the house told me (translation).

This example is typical of the processing of most Indic languages. P_2 (denoting parts) shows the constituents of the word strings. This needs sophisticated word segmenters and morphology analyzers.

State of the Art and Achievements

Despite the aforementioned challenges, the Indic language computing community has taken notable strides forward. This is seen on multiple fronts, such as corpus creation, NLP tool-building, end-user application development, research funding, collaboration, and standards and policy setting.

Fortunately for NLP, huge amounts of text in electronic form have become available in many walks of life (such as customer interactions in banks, reviews of online companies, judicial documents, contracts, e-books, and so on), paving the way for researchers to think about and apply powerful machine learning techniques to language technology problems. A case in point is the use of Europarl Parallel Corpus

There is no doubt that speech and natural language processing of Indic languages is hugely important and relevant, and has the potential to influence the lives and activity of at least 20% of the world's population.

[f] These are diacritic marks.

[g] We use transliterated Roman script for universal readability: H1$_1$- sentence no. 1, which is in Hindi; G1$_1$- word for word translation of sentence no. 1 called gloss; T1$_1$- translation in English of sentence no 1.

> **The Si-Ta translation system was developed as a solution to the scarcity of Sinhala-Tamil translators in the government sector. The system has already shown better performance than the commonly used Google Translate for the selected domain.**

in creating automatic MT systems. A game-changer came in 2005, when 110 pairs of statistical machine translation (SMT) systems were created by applying machine learning on this resource,[5] ushering in the era of SMT. Another paradigm shift came in the form of neural machine translation (NMT) in 2014, beating SMT by a wide margin.[1] The lesson is obvious: feed language data to ML algorithms to create NLP systems.

One of the authors of this article replicated the SMT and NMT research on Indian languages with his research team and wound up with state-of-the-art results for translation involving Indian languages and English.[6,9] The data used for training was the ILCI corpora[4] created at the initiative of the Technology Development in Indian Languages (TDIL) program of the Ministry of Electronics and Information Technology (MEITY), along with the Indian Institute of Technology Bombay (IIT Bombay) parallel corpus[8] created at the Center for Indian Language Technology of IIT Bombay.[h]

There have also been some isolated efforts to develop NLP applications to cater to specific needs in the region. One example is the Si-Ta machine translation system developed for Sinhala-Tamil to be used by the government sector of Sri Lanka. This translation system was developed as a solution to the scarcity of Sinhala-Tamil translators in the government sector. Despite the small parallel corpus used, the system has already shown better performance than the commonly used Google Translate for the selected domain.[10]

TDIL-MEITY has provided great service to the cause of Indian language technology (ILT) development. Since 2000, TDIL has been instrumental in initiating, funding, and sustaining research and development in ILT, including unicode standard, scripts, input methods, speech (http://www.iitm.ac.in/donlab/tts/), optical character recognition (OCR), MT, and cross-lingual information retrieval in Indian languages.[i] These initiatives have produced know-how, tools, and resources (like Indian-language Wordnets[j]) that are now ready to be commercialized through industry adoption and start-ups.

A recent initiative by NITI-Aayog,[k] the premier policy think tank of the Government of India, under the chairmanship of the Prime Minister of India providing both directional and policy inputs, brought together Indian academia, start-ups, industry, and research labs to discuss traction and monetization of ILT. It was decided to create an NLP access repository that would enable start-ups and industry to create large ILT applications, such as online review sentiment analyzers in Indian languages. The access repository will provide a platform from which to launch large applications.

The Bureau of Indian Standards of India's Ministry of Commerce recently set up a panel on Artificial Intelligence Standardization (LITD30).[l] This is the Indian mirror of SC 42, the sectional committee of the International Standards Organization (ISO) for AI standardization. Language Technology and its standardization is an important focus of LITD30, especially in the context of trustworthiness and certification (that is, automatic detection of fake news). Other noteworthy efforts on the subcontinent have been reported by the Language Technology Research Laboratory of Sri Lanka's University of Colombo,[m] the National Language Processing Centre of Sri Lanka's University of Moratuwa,[n] and the Center for Language Engineering[o] of Pakistan's Al-Khawarizmi Institute of Computer Science University of Engineering and Technology.

Way Forward

We close this discussion with a few pointers for moving forward:

- Although languages are quite distinct, there are also a number of similarities, in that all the languages can be represented by a superset of sounds, which is much less than the number of graphemes that make up all the languages. A unified representation is the current need to enable speech

[h] http://www.cfilt.iitb.ac.in

[i] Very informative articles on large consortia projects in ILT can be found at http://tdil.meity.gov.in/Publications/Vishwabharatnew.aspx.

[j] http://www.cfilt.iitb.ac.in/indowordnet/

[k] http://www.niti.gov.in/

[l] https://bis.gov.in/wp-content/uploads/2018/11/agenda-compo-litd-30.pdf

[m] http://ltrl.ucsc.lk/

[n] https://www.mrt.ac.lk/web/nlp

[o] http://www.cle.org.pk/

and language technologies. This will help pool low resources across various languages to build robust ASR systems for Indian languages.

- In the context of TTS, the major issue to be addressed is the input method. Text is available in multiple Indian scripts, but digital resources in terms of high-quality parallel corpora are few and far between. In the context of both ASR and TTS, generic acoustic models across various languages, generic language models in the former, and a generic Indic voice in the latter need to be designed. This will also address the issue of code switching.
- In TTS, code mixing must find ways to preserve the speaker's voice across languages. Further, the influence of the native tongue on a non-native tongue must be preserved. For instance, there are as many varieties of English as there are native tongues. Replacing non-native English (which is syllable-timed) with stress-timed English can make it difficult for the listener to understand.
- Text in social media generally includes code switching/mixing. Further, there are many words that have a local cultural connotation. Building language resources to address these requires the expertise of linguists, speech scientists, natural language processing engineers, and ethnographers.
- Data is the new oil, and NLP and ILT is no exception. There is no doubt that resources with quality and coverage need to be created, and created fast. Thinking creatively on how to engage even a small portion of 1 billion hands for resource creation is a must. Crowdsourcing, in spite of its criticism with respect to quality, seems to be the way forward. Providing attractive, helpful interfaces and remuneration can go a long way toward resource creation. In this context, the Language Data Consortium for Indian Languages (LDC-IL)[p] initiative of Central Institute of Indian Languages (CIIL) is noteworthy.
- Evaluation is the key to actual use of language resources and should be taken very seriously. Like TREC[q] (USA), CLEF[r] (Europe), and NTCIR[s] (CJK countries), India's Forum for Information Retrieval Evaluation (FIRE) initiative[t] has taken up the cause of evaluation in information retrieval and allied tasks. A FIRE-like initiative is needed for all areas of ILT.

Conclusion

Indic Language Computing (ILC) is too important a problem to be lying in oblivion. Given spectacular advancements to date in computing science and technology, Internet, AI, machine learning, and NLP, the time is ripe for a concerted thrust for realization and social penetration of ILC. The energy of the start-up echo system has to be harnessed with government support, and guidance from academia. Language resource creation is a precondition for ILC revolution, and as in all cases of large infrastructure building (roads, internet, gas lines, waterways), government sponsorship is needed for resource building.

Code mixing must find ways to preserve the speaker's voice across languages. Further, the influence of the native tongue on a non-native tongue must be preserved.

[p] http://www.ldcil.org/
[q] https://trec.nist.gov/
[r] http://www.clef-initiative.eu/
[s] http://research.nii.ac.jp/ntcir/index-en.html
[t] http://fire.irsi.res.in/fire/2019/home

References

1. Bahdanau, D., Cho, K. and Bengio, Y. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.
2. Bhattacharyya, P. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilinguality. *CSI J. Computer Science and Engineering 1*, 2 (2012).
3. Chakrabarti, D., Mandalia, H., Priya, R., Sarma, V., and Bhattacharyya, P. Hindi compound verbs and their automatic extraction. In *Proceedings of Computational Linguistics*, Manchester, U.K., Aug. 2008.
4. Jha, G.N. The TDIL program and the Indian language corpora initiative. In *Proceedings of LREC*, 2010.
5. Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Machine Translation Summit*, 2005.
6. Kunchukuttan, A., Mishra, A., Chatterjee, R., Shah, R. and Bhattacharyya, P. Shata-Anuvadak: Tackling multiway translation of Indian languages. In *Proceedings of the Language Resources Evaluation Conference*, 2014.
7. Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of LREC*, (Miyazaki, Japan, May 7–12, 2018).
8. Liu. B. *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers, 2012.
9. Murthy, R., Kunchukuttan, A., and Bhattacharyya, P. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. In *Proceedings of LREC*, 2019.
10. Ranathunga, S., Farhath, F., Thayasivam, U., Jayasena, S., and Dias, G. Si-Ta: Machine translation of Sinhala and Tamil official documents. In *Proceedings of the National Information Technology Conference*, 2019.
11. Subbarao K.V. *South Asian Languages—A Syntactic Typology*. Cambridge, 2012.

Pushpak Bhattacharyya (pb@cse.iitb.ac.in) is a professor in the computer science and engineering department of IIT Bombay, and director of IIT Patna.

Hema Murthy (hema@cse.iitm.ac.in) is a professor in the computer science and engineering department of IIT Madras.

Surangika Ranathunga (surangika@cse.mrt.ac.lk) is a senior lecturer in the department of computer science and engineering and a member of the faculty of engineering at the University of Moratuwa.

Ranjiva Munasinghe (ranjiva@mindlanka.org) is chief executive officer of MIND Analytics and Management in Colombo, Sri Lanka.

DOI:10.1145/3355625

BY VIVEK RAGHAVAN, SANJAY JAIN, AND PRAMOD VARMA

India Stack—Digital Infrastructure as Public Good

INDIA IS HOME to almost one-fifth of the world's population. Its scale and diversity rival those of continents, not countries. India has "official" 22 languages but unofficially 19,500 languages have been recognized as having 10,000 speakers or more.[a] There is incredible diversity, but also incredible disparity. About 45 million people still live in extreme poverty,[b] and less than 4% of its 1.3 billion people paid any income tax at all.[c]

At the same time, digital inclusion in India has taken off in a significant way in the last few years. It has 1.2 billion mobile connections and over 500 million Internet users.[d] India is now the world's second-largest market for smartphones, with an estimated 400 million smartphones in India having access to one of the cheapest mobile data plans in the world. India is incredibly young—about 50% of its population is below the age of 25, with approximately 65% of the population below the age of 35.[e] India expects to have 100 million people entering the workforce over the next 10 years. In short, the country is young, ambitious, and connected.

Social welfare is delivered through a complex network of over 950 schemes and funds by the Union government alone. The Union government spent close to $45 billion on subsidies last year. The states would cumulatively spend another $10 billion. A migrant worker population of over 453 million people,[f] moves from their homes either seasonally or permanently, adding to the complexity of welfare service delivery.

In 1985, the Prime Minister of India said that out of every rupee spent by the central government, only 15 paise (15%) reaches the beneficiary. This is because, distribution of welfare has typically taken place in kind, through a multi-layered supply chain. Realistically, it is estimated that leakages in welfare programs spanned from 10% to 60%, depending on the program.

Moreover, price subsidies tend to be regressive because they are untargeted. As the economic survey of 2015 defined it, "a rich household benefits more from the subsidy than a poor household."[g] The report found the bottom 50 of the country consumed less than 25% of the subsidized LPG (cooking gas). Similarly, 41% of the kerosene supplied through the public distribution system was lost to "leakages," and only 46% of the remainder went to poor households.

It is important to remember that some of these problems and numbers are as recent as 2015. Clearly, the state needed to move away from the price subsidy model to a more targeted

[a] Census of India, 2011; http://bit.ly/2Sysodk

[b] World Poverty Clock Statistics on India https://worldpoverty.io/

[c] Two crore Indians file returns but pay zero income tax. *Economic Times*, Oct. 23, 2018; http://bit.ly/2wWziiU

[d] IAMAI I-CUBE 2019 Report, http://bit.ly/2MQELCF

[e] Age structure and marital status, Census of India, 2011.

[f] Census 2011 Data; http://bit.ly/2Y9nrdj

[g] Economic Survey of India, 2014–15, Chapter 3: Wiping every tear from every eye.

A young woman applies for an Aarhaar card, the world's largest biometrics ID system.

PHOTO BY MELTING SPOT/SHUTTERSTOCK.COM

and efficient service delivery model. Starting in 2009, India began to create digital infrastructure to move from people and paper-intensive inefficient service delivery, to an efficient, direct, digital service delivery.

This was not just the need of the State, the Indian markets felt the same way. Despite its large size, and consistently high growth rates, the Indian markets have not turned out to be stellar for many players. The high cost of customer acquisition, KYC (Know-Your-Customer) process, various claims verification, and overall cost of business meant market players could not provide affordable and accessible products or services. A large population was not in the formal economy. This is the context that—beginning in 2009—over the next 10 years led to the creation of the India Stack.

India Stack

Leapfrogs. There have been various technologies that have played the role of infrastructure. While the technologies themselves are commendable, their real "disruptive" power has been what applications they enable. For example, the Internet may have been born of a specific need, but its success is because of its design. It was a mass-scale, open, and interoperable protocol. The use cases for the Internet were not restricted by the imagination of its founders.

The India Stack is a name given to a family of APIs, open standards, and infrastructure components that allow a user in India to demand services digitally. As of 2019, the services the India Stack offers are proving identity, completing KYC, making digital payments, signing documents digitally and sharing of data. While the list of APIs is growing, the APIs listed in Table 1 are now mature, well understood, and enable efficient delivery of services in India.

Why India Stack? Just like the modern Web, the India Stack did not come out of one place, but through multiple efforts by multiple teams. Each API or standard may have an owner and their own licensing nuances. It is a set of loosely coupled technologies and protocols, and there is no master directive. Each technology tries to do one thing and do it well. The innovation comes from the combinatorial use of these technologies by entrepreneurs and governments alike.

What they do have in common is that each lowers the cost of doing transactions. The reason for cost savings is multifold—it eliminates paper, but also eliminates the need for physical presence during a transaction. Digital payments eliminate cash and the cost of cash handling. It can also simplify compliance, such as in the case of KYC compliance for financial or telecom institutions. It could reduce "leakages" through the verification of identity and elimination of duplicates.

The breadth of India Stack and its potential use cases are too wide to

Table 1. India Stack's APIs.

Layer	Provider	APIs / Functionality	Uses
Presenceless	UIDAI	Authentication	Service Delivery Authentication Direct Benefits Transfer
Paperless	UIDAI	KYC	Bank Account Opening, SIM issuance
	CAs	eSign / Digital Signature	Contracts, Agreements
	Meity / Digilocker	Document	Consented Document Sharing
Cashless	NPCI / UPI	Payments	Retail payments, including P2P, P2M, Govt. through mobile
	AEPS, Aadhaar Pay	Payments	Cash deposit/Withdrawal, Transfers, Merchant payments using biometric auth
	IMPS	Payments	Remittances, Mobile payments
Consent	NBFC-AA	Financial Data	Personal Finance Management, Loan processing

Table 2. India Stack's impact factors.

Layer	Provider	APIs / Functionality	Volume / Impact
Presenceless	UIDAI	Authentication	1.2 Billion Enrolled 30.6 B Authentications to date, 745M in May 2019
Paperless	UIDAI	KYC	7.2 B eKYC to date 41.5M in May 2019
	Meity / Digilocker	Document	3.5 B digital documents
Cashless	NPCI / UPI	Payments	733M Transaction in May 2019
	AEPS, Aadhaar Pay	Payments	185M Transactions in Mar 2019

Source: Websites of various providers.

cover in depth here. We will focus on two of the components that are currently doing greater than 800 million transactions per month: Identity and payments.

The various components of the India Stack are at different levels of maturity. Table 2 illustrates some of the metrics for a selected subset of the systems.

Identity

In 2009, the Government of India undertook a program to give each resident of India an identity card. It was estimated that approximately 400 million people in India did not have an individual identity document.[h] The importance of identity for development is well understood. In India, this program was called Aadhaar, which translates to "foundation" in many India languages.

Where India differed from other similar programs of the time, the stated intent was to issue a secure, digital identity and not simply an ID card. The Aadhaar program scheme was presented as designed to be minimally intrusive, with the focus of the program on empowering every resident in two important ways. The first was to manage their identity; the second was to use their identity to prove who they are. The following sections, as well as Figure 1, help to explain the design.

Managing identity. The scale of the Aadhaar project and the diversity of India meant every assumption about a user's context, ability, or access to infrastructure would be challenged in the field. The design of the Aadhaar system preempted some of these challenges through simple design principles.

The first principle was to keep the data collected minimal. The Aadhaar system only collected four mandatory demographic variables: Name, address, gender, and date of birth, along with two voluntary attributes, namely, mobile number and email address. The voluntary attributes helped users manage their identity themselves online.

The choices around data collection, access controls, and system architecture should enforce hard limits to what is possible to minimize risk by design. In case of Aadhaar, biometric data cannot leave the Central Identity Data Repository of the Aadhaar in any circumstance. The feature is simply not present in the system, minimizing the likelihood of leaks whether accidental or intentional. These were part of Aadhaar's privacy by design principle.

The Aadhaar project was meant to provide an inclusive identity. No one should be left wanting an Aadhaar for lack of documentation or ability to register biometrics. Even if a resident could not furnish an existing identity document or an address proof to verify their details, a letter of introduction from their local representative would do. Similarly, there were exception processes for those with ailments or conditions that prevented them from successfully enrolling their biometrics. Inclusion in authentication was achieved through the availability of multiple factors of authentication including fingerprints, face, iris, and OTP.

The Aadhaar project implemented an ecosystem approach for solving problems of scale. For example, using standardized software, private enrollment operators were enlisted to go out and enroll citizens. They were paid by the Aadhaar project on a per successful enrollment basis. The enrollment data was end-to-end encrypted and deduplicated at the CIDR only. This lead to a rapid onboarding of users, reaching one billion enrollments in 5.5 years after launch.[i]

Projects such as Aadhaar, and components of the India Stack, are considered national assets that might outlive the existing vendor base. Propriety solutions offer short-term relief but may have a larger total cost of ownership. Using standardization and an

[h] Massive biometric project gives millions of Indians an ID. WIRED; www.wired.com/2011/08/ff_indiaid/

[i] Aadhaar Dashboard, UIDAI; https://uidai.gov.in/aadhaar_dashboard/

open architecture, the Aadhaar project was able to develop a vibrant and open vendor base for critical components of the hardware and software running Aadhaar. The project is deployed on commodity computing resources to prevent costly maintenance bills. Further, scaling to hundreds of millions of transactions per month and billions of records has not been a problem. This has led to massive cost savings, with each enrollment ultimately costing less than $1 per successful enrollment and authentications to approximately one cent per authentication.[j]

Using identity. Aadhaar is a digital identity, and its value is derived from the fact that to confirm the user who furnished the ID is indeed the true owner of that identity Aadhaar provides multiple channels for authentication. This allows governments and businesses to trust the person they are transacting with is truly who they claim to be.

Aadhaar serves as foundational identity and does not collect information on purpose of authentication. It has been envisioned that many domain-specific federated identities will be derived from Aadhaar. For example, India's tax ID—the Permanent Account Number (PAN)—uses Aadhaar to deduplicate its registers. Since these two databases remain separate, the CIDR has no information on the tax IDs of its users. This principle is also reflected in the institutional design of the program—the Unique Identity Authority of India (UIDAI)—which is a separate agency that does not fall under an existing function-specific ministry.

Aadhaar was aware of the growing privacy risks if identity and transaction data is collected in one central place. Hence, Aadhaar envisioned a federated model during use of Aadhaar.

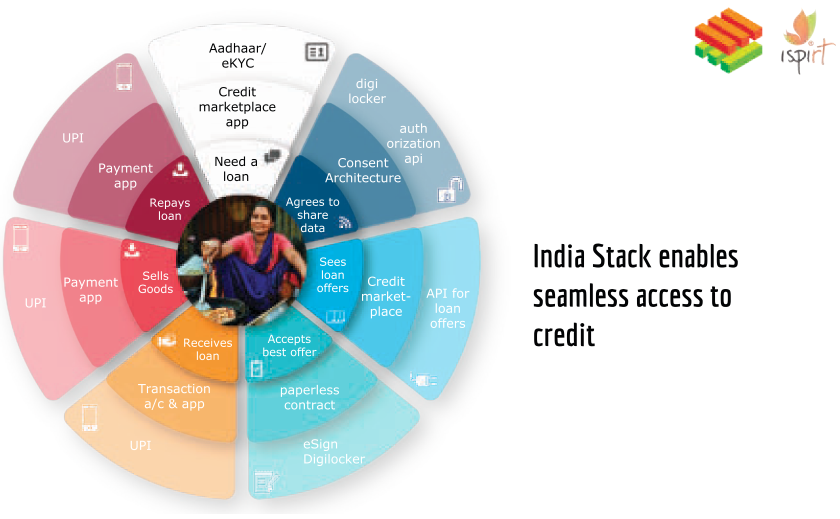
Payments

Despite having credit cards for more than 40 years, their penetration in India has been very low. In 2015, there were only approximately 20 million credit cards in the country and two million digital payment acceptance points for India's 1,300 million people,[k] indicating many features of card-based payment systems (for example, high cost of payments and cumbersome user experiences) were not effective at reaching most of the Indian market. The National Payments Corporation of India (NPCI) realized that for digital payments to be successful in India, it needed a low-cost payments system that worked for high volumes of low-value transactions.

The outcome of NPCI's deliberations was the Unified Payments Interface (UPI). The Unified Payments Interface is a protocol that simplifies the sending and receiving of value from any stored-value account to any other stored-value account. That is, the UPI specifications allowed sending money from bank accounts to bank accounts, but also from bank accounts to mobile wallets and loyalty accounts, among others.

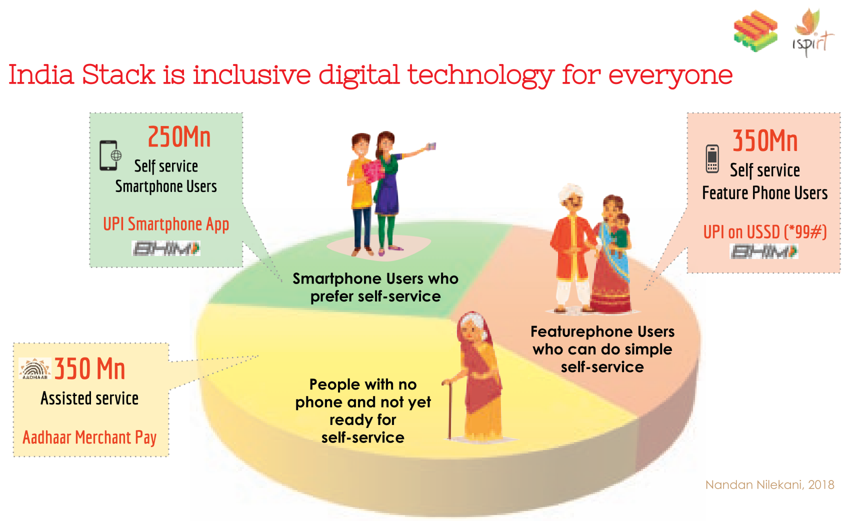
UPI provides a set of interoperable APIs that innovators use to build payment apps or make payments as a feature into their current workflows. Normally, this would have required bilateral agreements with all banks, but since almost all the banks in India use the UPI specifications for transferring money between bank accounts

Figure 1. India Stack's design layers.



India Stack enables seamless access to credit

Figure 2. India Stack's open architecture.


India Stack is inclusive digital technology for everyone

Nandan Nilekani, 2018

[j] Based on cost of Aadhaar project from A Cost-Benefit Analysis of Aadhaar. National Institute of Public Finance and Policy, Nov. 2012; http://planningcommission.nic.in/reports/genrep/rep_uid_cba_paper.pdf

[k] Bankwise CARD Statistics, RBI; https://rbi.org.in/Scripts/ATMView.aspx

and from bank accounts to mobile wallets and loyalty accounts, building a Venmo-like product in India is much easier. This is the reason India has seen an explosion of payment apps recently, including global players such as Samsung, Google, and Whatsapp.

How UPI did this was by first defining the Payments Markup Language. It standardized the instruction for push (sending) and pull (requesting) of money. All transactions are available on API endpoints, so that payments become a feature, not just an app. By standardizing and defining the Payment Markup Language, UPI could introduce features such as recurring payments that were previously only available although credit cards and tedious bank mandates.

Further, as part of its open architecture, UPI uses a pluggable authentication model, so that it is not dependent on any particular identity or mode of authenticating. This was important from the point of view of inclusion. In India, enabling digital payments cannot assume the presence of a smartphone. We were able to create two important apps on top of UPI to serve even those without smartphones. The first was the USSD based *99#, that enabled all transactions that a UPI app could do, but on a feature phone. The second was Aadhaar Merchant Pay. Using Aadhaar authentication, NPCI could transfer money from a user's bank account to that of a merchant without the need of a smartphone by the user. The consent to transfer is instead collected via biometrics at an agent's terminal who may have a smartphone or specialized point-of-sale machine.

UPI unbundled the "address" of payments. Instead of requiring users to remember an arbitrary combination of account numbers and routing numbers, UPI standardized the payment address. In UPI, every payment address is of the form "name@entity." This address is then resolved internally by NPCI to the correct account. Every account may have multiple payment addresses linked to it, so that the user may give john-banker@citi to his colleagues and john-gamer@sbi to his friends and both route money to the same underlying account from ICICI.

Figure 2 also alludes to the four-

DOI:10.1145/3353770

Privacy Concerns with Aadhaar

BY SUBHASHIS BANERJEE AND SUBODH SHARMA

The debate engendered by the Aadhaar project has propelled India from being a predominantly pre-privacy society to one in which privacy protection in digital databases has emerged as a major national concern. The welcome and scholarly Supreme Court judgment[8] has upheld privacy as a fundamental right, and informational self-determination and the autonomy of an individual in controlling usage of personal data have emerged as central themes across the judgment. The main privacy concerns with Aadhaar are:[1]

▸ *Identity theft.* Aadhaar is vulnerable to illegal harvesting of biometrics and identity frauds because biometrics are not secret information.[4,11] Moreover, possible leakage of biometric and demographic data, either from the central Aadhaar repository or from a point-of-sale or an enrollment device, adds to the risk.

▸ *Identification without consent using Aadhaar data.* There may be unauthorized use of biometrics to identify people illegally. Such violations may include identifying people by inappropriate matching of fingerprint or iris scans, or facial photographs stored in the Aadhaar database, or using the demographic data to identify people without their consent and beyond legal provisions.

▸ *Correlation of identities across domains.* It may become possible to track an individual's activities across multiple domains of service using their global Aadhaar IDs, which are valid across these domains. This would lead to identification without consent.

▸ *Illegal tracking of individuals.* Individuals may be tracked or put under surveillance without proper authorization or legal sanction using the authentication and identification records and trails in the Aadhaar database, or in one or more authentication-requesting-agencies' databases. Such records may reveal information on location, time, and context of authentication and the services availed.

Also, Aadhaar does not record the purpose of authentication. Authentication without authorization and accounting puts users at serious risks of fraud because authentication or KYC meant for one purpose may be used for another.[6] Recording the purpose of authentication is crucial, even for offline use.[2] Privacy-by-design is not achieved by self-imposed blindness.

Lack of protection against insider threats and lack of virtual identities—which were retrofitted in a limited way[9]—raise some serious privacy concerns, and the absence of a clear data usage policy and regulatory oversight exacerbates the problem.[1] Without a robust consent and purpose limitation framework and a regulatory access control architecture, the privacy concerns will remain. The inadequate privacy safeguards can potentially give the government of the day unprecedented access to information and power over its citizens threatening civil liberty and democracy.[3,5,7]

The Supreme Court's three-pronged proportionality test for the constitutionality of Aadhaar was based on determination of a rational nexus between the objectives and the means, of necessity—implying that the adopted means are the least intrusive for the purpose—and of balancing of extents to which rights are infringed.[7] Although the majority judgment upheld the constitutionality of Aadhaar, it struck down most of its uses on privacy grounds and limited its scope to only disbursement of welfare and income tax. The dissenting minority judgment, however, found Aadhaar to be unconstitutional in its entirety. Moreover, the Supreme Court of Jamaica has also recently struck down its very similar Jamaican National Identification and Registration Act (NIRA) as unconstitutional by heavily relying upon and extensively citing the dissenting Aadhaar judgment.[10] Judicious design of a national identity system that is respectful of fundamental rights is still very much an open problem.

References

1. Agrawal, S., Banerjee, S. and Sharma, S. Privacy and security of Aadhaar: A computer science perspective. *Economic and Political Weekly 52*, 37 (2017), 16.
2. Banerjee, S. and Sharma, S.V. An offline alternative for Aadhaar-based biometric authentication, 2018; http://bit.ly/330m8jn
3. Drezé, J. The Aadhaar coup, 2016; http://bit.ly/2IfqQSe
4. Khaira, R. Rs 500, 10 minutes, and you have access to billion Aadhaar details. *Tribune India*, 2018; http://bit.ly/2wW5wdY
5. Khera, R. *Dissent on Aadhaar: Big Data Meets Big Brother.* Orient Black Swan, 2019.
6. PTI. UIDAI suspends Airtel, Airtel Payments Bank's e-KYC license over Aadhaar misuse, 2017; http://bit.ly/2IJnjdR
7. *Puttaswamy, KS and Another v Union of India.* Writ petition (Civil) No 494 of 2012. Supreme Court judgment dated Sept.26, 2018; https://indiankanoon.org/doc/127517806/
8. *Puttaswamy, KS v Union of India.* Writ petition (Civil) No 494 of 2012. Supreme Court judgment dated Aug. 24, 2017.
9. Sharma, S. (via P.V. Singh). Virtual ID is a good beginning; much more remains to be done, 2018; http://bit.ly/2YxDmp5
10. Supreme Court of Judicature of Jamaica. Justice Sykes, B. Justice Batts, D. and Justice Hamilton, L-P. Claim No. 2018HCV01788 between Julian J. Robinson and The Attorney General of Jamaica, 2019; http://bit.ly/31r3XTg
11. Viswanath, L. Four reasons you should worry about Aadhaar's use of biometrics, 2017; https://thewire.in/featured/real-problem-aadhaar-lies-biometrics

Subhashis Banerjee (suban@cse.iitd.ac.in) is a professor in the Department of Computer Science and Engineering at Indian Institute of Technology Delhi, India.

Subodh Sharma (svs@cse.iitd.ac.in) is an assistant professor in the Department of Computer Science and Engineering at Indian Institute of Technology Delhi, India.

party model that is so important to UPI's success. In UPI, the payment-address-issuing entity is not necessarily the same as the one providing the underlying bank account. This means a user can use any app to send or receive money directly from their bank account. They are no longer restricted to just the app provided by their banking service provider. This has increased competitiveness to acquire users, and as a result the responsiveness and performance of bank apps has improved dramatically since the launch of UPI.

With over 800 million transactions worth more than US$1.9 billion being transacted monthly after approximately two years,[l] the Unified Payments Interface (UPI) is the fastest-growing open-loop digital payments platform in the world.

Criticism and Evolution

As Aadhaar gained coverage, traction, and the trust of service providers as a unique and robust proof of identity, it began to be requested (and sometimes mandated) as a foundational document across a variety of public and private services, in particular, for government subsidies, banking, and telecommunications.

As a result, there was pushback from media, civil society, and academics around issues of privacy and security of individual data, and the possibility of exclusion from access to services due to lack of an Aadhaar or due to authentication errors. Meaningful engagement on all criticisms is not possible in this article, the issues are wide ranging and need detailed, nuanced discussions on design trade-offs.

What we would like to highlight is some of the outcomes from the critique of Aadhaar. The UIDAI was able to see the increasingly vocal demand for better privacy controls, resulting in design changes to the program as it evolved. Aadhaar has rolled out a number of features to further enhance the security, privacy, and inclusion of the Aadhaar system.

Biometric capture devices are registered with the Aadhaar ecosystem and all biometrics captured are signed and encrypted at the capture device to prevent replay attacks. Residents can lock and unlock (for short periods of time) their biometrics using the multiple channels such as the Aadhaar mobile application or the Web portal.

Aadhaar introduced temporary virtual IDs that allowed users to mask their Aadhaar numbers during an authentication request. The means the Aadhaar number does not need to be shared with an authenticating agency. In the digitally signed response, Aadhaar returns agency-specific UID tokens, which are unique and cannot be correlated across agencies. In addition, residents can lock their Aadhaar number and authenticate using *only* the virtual ID.

Aadhaar has introduced the concept of offline KYC verification, which allows residents to directly share their digitally signed KYC information with a verification agency XML/QR code formats. This allows residents to share non-tamperable credentials without direct involvement of the Aadhaar system. Local validation of the photograph through face matching and mobile number are possible. Sensitive data such mobile number is stored using a one-way hash; the data is revealed only if residents share the data with the verification agency.

Problems with authentication using fingerprints by manual laborers or senior citizens were addressed through the introduction of multiple biometric modalities such as face and iris matching. In addition, multiple modalities can be combined through fusion to further reduce rejections in the field. Finally, exception processes are put in place to ensure 100% of residents can authenticate using the Aadhaar system. Aadhaar's open architecture meant such a solution could be rolled out quickly in response to public demand.

The criticism and civil society movement also bought into the public discourse India's lack of a Data Privacy Law, which is necessary whether or not there is an Aadhaar. While trying the Aadhaar case, the judges were forced to ask if the constitution guarantees a fundamental right to privacy. A nine-judge bench found the answer was affirmative.[m]

A second Supreme Court judgment declared Aadhaar did not intrinsically violate an individual's fundamental right to privacy, but its mandated use ought to be restricted only to government-provided subsidies and benefits, tax collection, and other proportional use cases where permitted by law.[n]

While it may seem contentious and politically charged, such conversations are a feature, not a bug, of democracy. The executive, judiciary, and UIDAI were responsive to the public's needs and evolved the system based on what the people wanted. Our experience underscores the importance of stakeholder conversations during the design and implementation of the program.

Conclusion

India's experience with creating digital infrastructure platforms as public goods offers multiple lessons learned in technology, system, and regulatory architecture. It demonstrates how multiple such systems can be leveraged in concert—such as the India Stack—for development objectives. Governments and businesses alike are building for diverse use cases on top of the stack. By lowering the transaction costs of serving the poor, we are achieving better inclusion.

Such digital infrastructure is not a unique requirement in India. It is estimated that approximately 161 countries currently have or are building their own digital ID systems. Many countries have local interbank payment systems and are now looking to upgrade them for a mobile-first world. As various countries build their own systems, the Indian experience with Aadhar serves as a real-world example to learn from. Even if the systems may look different, we believe the principles adopted in their development would serve well globally.

[l] UPI Product Statistics; https://www.npci.org.in/product-statistics/upi-product-statistics

[m] *Justice K.S. Puttaswamy (Retd) vs Union of India*, Aug. 24, 2017.

[n] *Justice K.S. Puttaswamy (Retd) vs Union of India*, Aug 26, 2018.

Vivek Raghavan (vivek.raghavan68@gmail.com) is Chief Product Officer of UIDAI, Bangalore, India.

Sanjay Jain (snjyjn@gmail.com) is Chief Innovation Officer of CIIE, IIMA, Bangalore, India.

Pramod Varma (pramodkvarma@gmail.com) is Chief Architect at UIDAI, Bangalore, India.

DOI:10.1145/3344432

BY CHARLES ASSISI, AVINASH RAGHAVA, AND NS RAMNATH

The Rise of the Indian Start-Up Ecosystem

WALK INTO ANY one of the many start-up events organized across India, and inevitably the image of an Indian bazaar comes to mind: people rushing around, shouting, bargaining, answering phones with great excitement, laughing loudly, boasting, blushing, and generally being optimistic, as if they are at the beginning of a rising trend of well-being.

Such optimism might seem justified. According to data compiled by *Fortune* magazine,[a] from just eight 'unicorns' in 2015, the number of start-ups in India valued at more than $1 billion has grown to 26. What is interesting is that in 2018 alone, India added eight unicorns to the club.

These include diverse entities such as Ola, started in India as a competitor to Uber and has since expanded its footprint into the U.K. (and is eyeing Australia); an insurance aggregator called PolicyBazaar; the e-commerce site Paytm Mall; an eyewear retailer called Lenskart; food technology aggregators such as Swiggy and Zomato, and hotel-room aggregators like OYO and FabHotels.

Thousands of entrepreneurs start up every year and aspire to become one of the new unicorns. Venture capitalists invested over $20 billion on start-ups last year, and evidence suggests they are likely to invest much more by the end of this year.

The rise of the Indian start-up ecosystem can be characterized by three major changes over the last decade:

1. A shift from models copy-pasted from elsewhere to the creation of models built for India.
2. A move from the IT services model to technology products.
3. A statement of intent from entrepreneurs that the time for *Jugaad* is over, and cutting-edge innovations are where the future lies. The Hindi word *Jugaad* roughly translates as "to work around." The notion was a result of resource constraints faced by a number of enterprising Indians, especially those living in rural areas. *Jugaad* is a well-researched theme and has been extensively documented by Navi Radjou,[b] a French-American scholar based in Silicon Valley, in his book *Frugal Innovation*.

These changes offer interesting insights for the start-up ecosystem in India and across the world.

From 'Copy-Paste' Models to Local Innovation

The current crop of Indian start-ups trace their origins to the mid-1990s and late 2000s. They were driven by entrepreneurs and venture capitalists (VCs) from the U.S. (Silicon Valley, in particular), or were heavily

[a] http://bit.ly/2IHMrl3

[b] https://thinkers50.com/biographies/navi-radjou/

influenced by what was happening there. Many were engineers who studied and worked in the U.S. and got to witness firsthand the impact start-ups can have on an ecosystem. When the dot-com bust happened in the early 2000s, the start-up ecosystem picked itself up, and companies such as Google and Amazon not only survived, but thrived.

That being the case, they reasoned, the economy in India holds much promise—driven by a growing middle class, a demographic dividend from a huge population of working age, and above all, by government policies that seemed to be growing friendlier for businesses.

Technology—computers, Internet, software, devices—was going global as well. If Amazon can sell books to Americans, they reasoned, an Indian version of Amazon can do the same for Indians. Thus, by the end of the 2000s, India had a bunch of start-ups that drew their optimism mainly from the success of American tech companies.

For example, Flipkart, an e-commerce company which was bought by Walmart in 2018 for $16 billion, was started by two engineers—Sachin and Binny Bansal, who had worked for Amazon in India. As Amazon did back in the 1990s, the Bansals started in 2007 by selling books. They were not the first to do it; there were a half-dozen others already selling books in India, but the Bansals understood the importance of the customer experience as few others did.

The rather tepid performance of their predecessors did not deter them, because by 2008, Internet penetration was higher, broadband was picking up, and costs were coming down. They got support from investors including Accel, and were focused on customer satisfaction, a mantra they were initiated into at Amazon.

However, they soon realized that to satisfy customers, which boiled

PHOTOS BY: (TOP) PHUONG D. NGUYEN/SHUTTERSTOCK.COM; (MIDDLE) GRZEGORZ CZAPSKI/SHUTTERSTOCK.COM; (BOTTOM) JASON ARORA/SHUTTERSTOCK.COM

down to ensuring timely deliveries, they needed to get a grip on inventories of books with their own vendors, but they could not, because many of their vendors had not digitized their systems (and those that had were not connected to the Internet; if they could connect, they lacked compatible databases). The backend was not just about building your backend, but also pushing various stakeholders to build theirs.

Even on the customer side, the Bansals realized, while many were ready to place orders online, they were reluctant to make payments online. Many Indians either did not have credit cards, or the many who did have them were uncomfortable using them for online transactions. That got in the way of customer adoption.

Flipkart's answer to the problem was simple: pay cash on delivery. This simple tweak opened up latent demand, and was one of the reasons Flipkart grew quickly. The reluctance on the part of Flipkart's customers to transact online offered them a peek into inefficiencies in the payments space.

Yet as Haresh Chawla, a partner at the private equity firm True North, pointed out in an essay on FoundingFuel.com,[c] the Bansals could not capitalize on their early gains. This was happening as U.S. and Chinese entities were eyeing India, while other Indian entrepreneurs imagined new possibilities.

That is how digital wallet Paytm was created by Vijay Shekhar Sharma. Not only is that firm a unicorn now, it integrated backward to build a e-commerce portal called Paytm Mall to compete with Flipkart.

Many of the new possibilities had to do with the launch of the unified payments interface (UPI), a mobile platform that allowed customers to transfer money as simply as sending an SMS. UPI led to the large-scale entry of banks into the realm of payment apps.

This is not to suggest Flipkart did not put up a good fight. The first popular UPI app was from a start-up called PhonePe (which was started by former Flipkart engineers, and was eventually acquired by Flipkart). One of the most interesting PhonePe products was a point of sale (POS) terminal that cost a fraction of the price of traditional POS terminals, and that allowed credit cards to be swiped. While a standard POS terminal could cost Rs 20,000 (approximately US$300), retailers could get PhonePe's POS unit with a security deposit of less than Rs 700 (about US$10).

However, the most common "POS terminal" was just a QR code used by Paytm, whose "cost" is as low as that of printing a QR code.

The drastic reduction in costs, along with the targeting of specific niche markets such as vegetable vendors, roadside tea stalls, and generally people closer to the bottom of the economic pyramid, represented a big shift in the focus of Indian start-ups.

Most of them no longer had to look at the U.S. for inspiration. Instead, they were looking at problems faced by people who live in what investors now call 'India 2' and 'India 3', the lower levels of India's wealth pyramid. People such as Vijay Shekhar Sharma focused harder on this segment because they could see the economic potential there before the founders at Flipkart did.

The assumptions made for India 1 (those at the top of the pyramid and closer to U.S. markets) no longer apply to those who live in India 2 and India 3.

In 2015, former chief economic advisor to the Indian government Arvind Subramanian used phones as a proxy to separate the three segments; the approximately 200 million people who use smartphones, the 400 million or so who use feature phones, and those lacking access to any phones.

Management guru C.K. Prahalad had argued that there were fortunes to be made at the bottom of India's wealth pyramid. In addition, China had demonstrated that if you had a large population, you could build large businesses

Many Indians either did not have credit cards, or those who did were uncomfortable using them for online transactions. That got in the way of customer adoption. Flipkart's answer to the problem was simple: pay cash on delivery.

[c] http://www.foundingfuel.com/article/saving-private-flipkart/

and make interesting innovations based on domestic markets.

This realization led a range of start-ups that targeted specific segments across Indias 1, 2, and 3.

Banking and finance have always been early adopters of technology, and financial technology start-ups began trying to solve some of the problems that banks and financial institutions could not solve with the burden of brick-and-mortar infrastructure.

One of the early validations that using digital technologies could help in financial inclusion came from an experiment that IFMR Trust (now Dvara Trust) did with its KGFS model. For a long time, financial inclusion meant micro credit.

However, the designers of the Kshetriya Gramin Financial Services (KGFS) model, which included Nachiket Mor, then heading up the ICICI Foundation, and Bindu Anant, who was with venture capital firm IFMR Trust, wanted to create a system that didn't just give credit, but also provided a range of financial services, including savings and insurance products. Savings products were not yet being offered to the poor, because the money they put in did not even cover the paperwork needed to accept it. As an experiment, the KGFS designers digitized the entire process in money market mutual funds, and found it worked.

While KGFS could not scale up some of its products, it showed fintech how going digital can help financial inclusion by bringing down transaction costs.

Fintech is one example of how Indian start-ups, even as they pursue growth and profits, also fill the gaps that government, businesses, and the social sector either would not or could not in the past. By making technology work, many start-ups today are aligned with broader societal goals.

As Nandan Nilekani, chairman of Infosys and former chairman of the Unique Identification Authority of India (UIDAI), said during a conversation with Microsoft CEO Satya Nadella in Bangalore: "When you think of the challenge of taking the country with $2,000 per capita to $20,000 per capita, it is really a major challenge. You have to fix basic things like health, education, and access to financial services. The classical way of doing that would have been to say 'let's have more doctors, let's have more teachers,' and so on. That's certainly not possible in the timeframe that we have. If you want to get everyone educated, and if you take 25 years to do it, then they will be adults and you can't do it. The only way to square the circle is by using AI (artificial intelligence) and the cloud to deliver personalized health, education, and finance to a billion people. That will drive the economy. For a country that has low per-capita income, to use this as a strategic tool … is very important."

Sachin (left) and Binny Bansal.

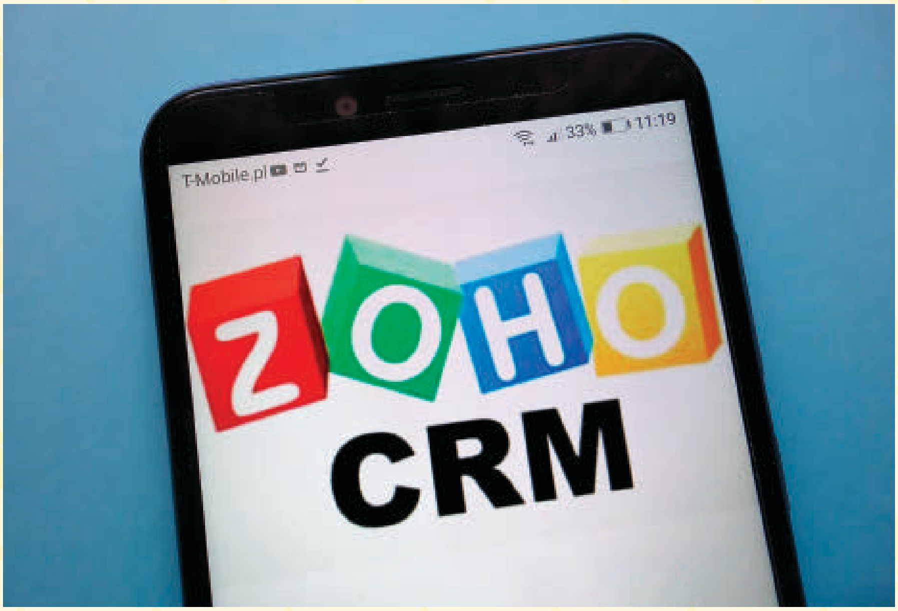
How India earns.

- 1 Indian earns 30% of the total and makes over Rs 1.5 lakh a month
- 14 Indians earn 30% of the total and make around Rs 20,000 a month each
- The next 30 Indians earn 30% of the total and make Rs 8,000 per month each
- The poorest 55 Indians earn 10% of the total and make only Rs 1,500 per month each
- 1 Indian owns 53% of the wealth
- 9 Indians own 23% of the wealth
- 40 Indians own 20% of the wealth
- 50 Indians at the bottom own only 4.1% of the total

India's unicorns.

Name	Value*	Incorporated	Industry	Investors
BigBasket	$1	6/5/19	E-commerce/ marketplace	Alibaba Group, Bessemer Venture Partners, Helion Venture Partners
Dream11	$1	9/4/19	Sports/gaming	Kaalari Capital, Tencent Holdings, Steadview Capital
Udaan	$1	3/9/18	E-commerce	DST Global, Lightspeed Venture Partners, Microsoft ScaleUp
PolicyBazaar	$1	6/25/2018	Fintech	Info Edge, Softbank Capital
InMobi	$1	2/12/14	Adtech	Kleiner Perkins Caufield & Byers, Softbank Corp., Sherpalo Ventures
Shopclues	$1.1	12/1/16	eCommerce/ Marketplace	Nexus Venture Partners, GIC Special Investments, Tiger Global Management
Swiggy	$3.3	6/21/2018	On-demand	Accel India, SAIF Partners, Norwest Venture Partners
Hike	$1.4	8/16/2016	Social	Foxconn, Tiger Global management, Tencent
Delhivery	$1.6	2/27/2019	Supply Chain and Logistics	Times Internet, Nexus Venture Partners, SoftBank Group
ReNew Power	$2	2/14/2017	Energy and Utilities	Goldman Sachs, JERA, Asian Development Bank
Zomato	$2.18	10/4/15	Social	Sequoia Capital, VY Capital
BYJU'S	$5	7/25/2017	Ed Tech	Tencent Holdings, Lightspeed India Partners, Sequoia Capital India
Oyo Rooms	$4.3	9/25/2018	Travel Tech	SoftBank Group, Sequoia Capital India, Lightspeed India Partners
Ola Cabs	$6.2	10/27/2014	On-demand	Accel Partners, SoftBank Group, Sequoia Capital
Snapdeal	$7	5/21/2014	E-commerce	SoftBankGroup, Blackrock, Alibaba Group
One97 Communications	$10	12/5/15	Fintech	Intel Capital, Sapphire Ventures, Alibaba Group

* In billions as of press time.



PHOTOS BY PIOTR SWAT/SHUTTERSTOCK.COM

From Services to Products

Before Nandan Nilekani became known for his work on Aadhaar, India's national identity program, the largest program in contemporary human history that created unique identification numbers for over a billion people in record time, he was best known as a co-founder of Infosys, one of the top software outsourcing firms in the country. He is also the person who gave the title to *The New York Times* columnist Thomas Friedman's best-selling book, *The World is Flat*, which argued that access to technology by people across the world has taken away the advantages advanced countries enjoyed during most of the industrial age.

In India, many entrepreneurs first came to understand the power of information technology by looking at the enormous success of companies such as Infosys, TCS, and Wipro. They are now multibillion dollar companies, employ hundreds of thousands of people, and have raised the standards of corporate governance in India.

However, success also comes with its disadvantages. Investors and entrepreneurs can become addicted to the metrics that work for IT services, but do not work for products. Due to huge labor arbitrage, IT services were operating on high margins. They could record revenues almost as soon as they started deploying resources, which was comforting to investors.

To create IT products, on the other hand, expenses are incurred up front, which also has its risks. After investing to create a product, if the product bombs, the investments sink. Fear of such a situation stopped Indian IT services companies that had huge cash balances and literally zero debt from investing in products.

There were some exceptions. Infosys read the writing on the wall and built a fairly successful core banking product called Finacle. Cognizant rolled out a three-horizon strategy, with its CEO Francisco D'Souza directly focusing on new businesses.

While their efforts were looked

down upon initially, it is only now that investors are beginning to appreciate why products matter in the long run. As the cost of labor started to rise over the years, margins from IT services started to come down, and competition emerged from other geographies where labor is cheaper.

Between entities such as Infosys, Cognizant, TCS, Wipro, and some mavericks who pushed the pedal to the metal on products, start-ups in India are looking at a contemporary narrative. One example of such a maverick is Zoho, which was founded in 1996. Zoho builds a range of Web-based technology tools aimed at improving the productivity of businesses. Its founder, Sridhar Vembu, on encountering the unfair conditions of venture capitalists, swore to build a company without going to VCs. Zoho's revenue is estimated to be around $500 million.

One interesting innovation of Zoho's is its training program, which selects students (typically from poor backgrounds) from schools and teaches them to code. Some students from early instances of the program are now product managers at Zoho.

Zoho is not the only enterprise tech company that has made its mark in the product space. Freshworks, started by a former employee of Zoho, also focuses on the SMB market and has innovated on the Inside Sales model.

India today has more than 10 B2B companies with Unicorn status; some reached this milestone in less than three years. Companies like BlackBuck, Udaan, Power2SME, Delhivery, and Capillary Technologies are trying to solve some of India's problems. Deep technology security companies like Druva, Qubole, and CloudCherry are leveraging India as a base for their development.

From *Jugaad* to Cutting Edge

For years, Indian innovation was mostly associated with the word *Jugaad*. The flip side to this is that it indicates short-term vision that moves from one kind of 'duct tape' to fix a problem to another, instead of asking deeper questions about design.

Yet *Jugaad* has evolved in its own way, and has come to mean frugal innovation; coming up with solutions using minimal resources, for a market that could not afford expensive products or solutions. Consumer products company Godrej developed ChotuKool, a portable refrigerator that consumes minimal power, for villages facing continuing power outages.

Underlying the change from *Jugaad* to frugal innovation is the belief that it is possible to build world-class products with limited resources. One of the insights of C.K. Prahalad is that innovations made for the bottom of the pyramid often work for segments that occupy the higher levels.

These three shifts—from the copy-paste model to local innovation, from software outsourcing to product development, from *Jugaad* to frugal innovation—have given way to a new breed of start-ups that are hugely aspirational.

Sharad Sharma, co-founder of the Indian Software Products Industry Round Table (iSPIRT), makes the distinction between mercenary start-ups whose primary goal is to make money, and missionary start-ups whose primary goal is to solve impossible problems.

An example of such a missionary start-up is TeamIndus, India's only entrant in the competition for the Lunar X Prize, in which teams are challenged to "land a robot on the surface of the Moon, travel 500 meters over the lunar surface, and send images and data back to the Earth." TeamIndus did not win the Lunar X Prize; no one did. But the point about start-ups such as TeamIndus is their goal is not just to win a prize, but to show it is possible to aim high.

TeamIndus is just one example of a start-up that many would not instinctively associate with India. There are many start-ups in India that work on cutting-edge technologies. Medical diagnostics start-up SigTuple Labs uses AI to analyze visual medical data, while Mitra Biotech is advancing personalized oncology treatment and supporting more effective and efficient cancer drug development. GreyOrange is a focused robotics warehouse management company. Julia Computing has developed a unique, high-performance programming language with rich applications in AI and machine learning capabilities.

Lessons

These three shifts offer two broad lessons to start-up ecosystems across the world:

The starting point does not matter; the direction does.

It does not matter where the story starts, or where the motivation comes from. It could have been in copy-pasting Western business models, bidding for software coding services based on labor arbitrage, or even *Jugaad*. What matters is the evolution.

Context matters; start-ups come to life in a society.

However, evolution seldom happens on its own. Evolution happens within a context; when entrepreneurs start solving a problem for the society in which they live, they experiment, scale up, and reach out to new customers.

Charles Assisi is co-founder and director of Founding Fuel, Mumbai, India. He is co-author (with NS Ramnath) of *The Aadhaar Effect: Why the World's Largest Identity Project Matters.*

Avinash Raghava is Community Platform Evangelist for Accel, Bangalore, India.

NS Ramnath is part of the founding team at Founding Fuel, Bengaluru, India, where he now serves as a senior writer.

The views expressed here are the authors' and do not necessarily reflect those of their employers.

DOI:10.1145/3343454

BY SUPRATIK CHAKRABORTY AND VASUDEVA VARMA

Highlights of Software R&D in India

INDIA IS A software superpower today. This achievement rests on more than four decades of work spanning software processes, rigorous engineering and value-adding technologies, among others. In this article, we present highlights of some of these activities. This regional section also contains other articles that complement this account of exciting work in software systems stemming from India.

The Indian software industry is currently valued at approximately US$180 billion, and is projected to touch $350 billion by 2025.[a] It serves most regions of the world and employs four million people directly and 13 million people indirectly.

Developing and delivering software solutions at this scale across diverse domains requires constant effort to improve the processes, tools, and platforms. Therefore, almost all major software companies in India have separate teams to address long-term research and development problems.

Research groups at Indian educational institutions such as Indian Institute of Science (IISc), and some of the Indian Institutes of Technology (IITs) and Indian/International Institutes of Information Technology (IIITs) have also contributed in significant measure to this success story. Besides technology transfers arising out of industry-academia collaborations, work originating in Indian labs have consistently appeared at top-tier conference venues like ICSE, POPL, PLDI, FSE, CAV, TACAS, SAS, ACM TOPLAS,

[a] https://www.ibef.org/industry/information-technology-india.aspx

The Bagmane Tech Park in Bengaluru, India, is a software technology office space and home to some of the biggest tech corporations worldwide.

PHOTO BY NOPPASIN WONGCHUM/SHUTTERSTOCK.COM

and IEEE TSE. In fact, ICSE 2014 was held in Hyderabad and POPL 2015 was held in Mumbai under the General Chairship of Pankaj Jalote and Sriram Rajamani, respectively.

India also has its flagship annual conference called Innovations in Software Engineering (ISEC), which provides a platform for sharing experiences of various research groups.

Indian Industry's Leadership in the Software Process

Up until the 1990s, the world of software was filled with stories of delayed and poor-quality software projects. For improving this scenario, a five-level Capability Maturity Model (CMM)[b] was developed by the Software Engineering Institute (SEI). CMM is a framework and model for evaluating and improving the software development process in an organization in a staged manner, and has been adopted by companies across the world. Indian software companies took a lead in deploying this model for improving software quality and productivity. In the early stages of model deployment, a large percentage of the companies at high maturity levels (CMM level 5) were from India—a situation that continues even today. Companies across the world, and countries desirous of developing their software sector, wanted to learn from the Indian experience in employing rigorous software processes using quantitative tech-

[b] https://en.wikipedia.org/wiki/Capability_Maturity_Model

Developing and delivering software solutions at this scale across diverse domains requires constant effort to improve the processes, tools, and platforms.

niques for managing them. Indian companies shared their experience in conferences, workshops, seminars, as well as through books like *CMM in Practice* and *Software Project Management in Practice* (both authored by Pankaj Jalote), which were translated in various languages such as Chinese, Japanese, Korean, and French.

Strides in Software Engineering

Software engineering is yet another pillar on which India's software success story rests. Research groups in Indian companies and universities today are exploring problems in several areas such as foundations of software engineering, quality assurance, architecture and design, security, software engineering for the cloud and mobile environments, software engineering education, and applying AI/ML in the software engineering domain. Here, we highlight a few of these prominent activities:

The Research & Innovation unit of Tata Consultancy Services (TCS) has been developing MasterCraft,[c] a toolset for supporting model-driven software development, for close to 17 years. MasterCraft is comprised of three major components. First, it has a set of meta-models to specify layers of a typical distributed architecture such as graphical user interfaces, services layer, and data manager layer, among others. Next, to facilitate smooth integration, MasterCraft provides component abstraction that helps view a software system as a set of interdependent components that can be specified, developed, and tested independently. Finally, MasterCraft incorporates a set of core technologies such as meta modeling, model editing, ensuring model well-formedness as well as internal consistency, and model-to-model and model-to-text transformation.

MasterCraft has made a huge business impact, delivering more than 70 large business applications across the world on a multitude of technology platforms and architectures. Its use has also led to 50+ top-tier publications and 20+ patents. Much of this research also found its way into international standards at Object Management Group, and has contributed to three core standards.

A comparably significant effort at Infosys has been the development of the Infosys DevOps Platform (IDP)[d] that helps organizations accelerate their agile and DevOps journey in quality, at scale, and at speed. It has ready-to-use pipelines for more than 25 technologies and prebuilt integration with over 70 open source or commercial tools. IDP is built on open source resources and is available as an open source project. It has made a huge impact in software development processes across the world in terms of its adoption—more than 100 projects in 30+ organizations worldwide, with more than 5,000 Infosys engineers trained and serving various clients.

In addition to these industry-led efforts, there are several exciting software engineering projects happening in Indian academic labs as well. Automated usability evaluation of mobile applications is one such project from IIIT Hyderabad. Usability is considered one of the primary factors for end users to adopt mobile devices/applications. The IIITH group's research led to the development of a code analysis-based usability evaluation framework for mobile apps that can be used at the predesign stage to enhance productivity or at the post-design stage to check conformance to specific usability guidelines. Automated evaluation of the mobile application is done using quantitative metrics and AI/ML-based methods.

Program Analysis and Verification

Going beyond CMM and software engineering, program analysis and formal verification are increasingly viewed as technologies that add value to enterprise and mission-critical software, both during its development and as an end product. This is particularly true for software that runs on potentially unreliable hardware and yet must provide guarantees of performance, security, functionality, and so on. Industrial and academic research groups in India have been consistently pushing the frontiers of program

[c] https://mastercraft.tcs.com/

[d] http://bit.ly/2XDNQiZ, page 12

analysis and verification, targeting both scalability and precision. Here, we highlight a few contributions:

Precise and scalable pointer analysis is known to enhance the quality of other program analyses by uncovering the indirect manipulation of data and indirect flow of control. However, it is challenging to scale an exhaustive flow- and context-sensitive pointer analysis to large programs in languages like C and C++. Most approaches begin with scalable but imprecise methods and try to increase their precision.

The programming languages research group at IIT Bombay has taken the opposite approach in that it began with a precise method and attempts to increase its scalability without compromising on precision and soundness. This has made it possible to strike a fine balance between precision and performance in pointer analysis, beyond what could have been achieved earlier.

Yet another area that has seen important contributions from India is static assertion checking. The primary challenge here is to reason about programs with loops, especially those that manipulate data structures like arrays and lists. Verifying asynchronous and concurrent programs presents yet another set of technical challenges related to races, deadlocks, memory consistency models, and the like. Research groups at IISc, IIT Bombay, TCS, and Microsoft Research India among others, have been working on abstraction and constraint solving-based techniques to handle complex assertion checking tasks like these. Their work has been reported at leading venues like CAV, TACAS, TOPLAS, and SAS, and some of these technologies have also been inducted in industry-scale tools. In the 2019 edition of the *Competition on Software Verification*, one of India's entries—VeriAbs from TCS—nabbed the top position in the "ReachSafety" category of the competition.

The P# project[e] from Microsoft Research India harnesses program analysis and formal verification techniques to develop a unique actor-based programming model for event-driven asynchronous applications and services. This model allows programming concurrent applications at a higher level of abstraction so the code more closely resembles its design. It is a natural fit for programming reactive distributed systems.

In addition to an efficient and lightweight runtime, P# provides the capabilities of writing detailed safety and liveness specifications. A testing engine controls the scheduling of the program, as well as all declared sources of nondeterminism (for example, failures and timeouts), to systematically explore behaviors, looking for violations of the specifications. If a bug is found, the testing engine reports a deterministic reproducible trace that can be replayed in the debugger. P# has been used by several teams in Microsoft Azure to write cloud services, who have reported dramatically increased productivity. Further, there have been nearly zero crashes reported for components designed using this model.

Sankie[f] is yet another project from Microsoft Research India where multiple technologies, including program analysis, root-cause analysis, and data-driven machine learning techniques are being harnessed together to improve the software development process.

The Road Ahead

India has historically played a leadership role in global software development. As newer technologies like data-driven techniques get embedded in software design, the underlying processes, engineering, and technologies will inevitably need to adapt. Indian companies and researchers are already gearing up for this. While the nature of problems to solve will change over the years, the Indian software R&D community appears sufficiently well grounded and equipped to rise up to the challenge.

> India has historically played a leadership role in global software development. As newer technologies get embedded in software design, the underlying processes, engineering, and technologies will inevitably need to adapt.

[e] http://bit.ly/2Xdotrf

[f] https://www.microsoft.com/en-us/research/project/sankie/

Supratik Chakraborty is a professor in the Department of Computer Science and Engineering at I.I.T. Bombay, Mumbai, Maharashtra, India.

Vasudeva Varma is a professor at I.I.I.T. Hyderabad, Telangana, India.

DOI:10.1145/3343449

BY MEENA MAHAJAN, MADHAVAN MUKUND, AND NITIN SAXENA

Research in Theoretical Computer Science

THEORETICAL COMPUTER SCIENCE has been a vibrant part of computing research in India for the past 30 years. India has always had a strong mathematical tradition. One could also argue that in the 1980s and 1990s, theory offered a unique opportunity to keep up with international research in computing despite limited access to state-of-the-art hardware.

The Annual International Conference Foundations of Software Technology and Theoretical Computer Science (FSTTCS) was launched in 1981. FSTTCS[2] allowed Indian researchers a natural opportunity to interact with leading academics worldwide.

Another early impetus was funding for international collaboration through agencies such as the Indo-French Centre for Promotion of Advanced Research (CEFIPRA).

Research Highlights

Algorithms. Maximizing the flow that can be routed in a network is one of the most well-studied algorithmic problems, with immense practical applicability. In the 1970s, when computer science research in India was taking root, Sachin Maheshwari and his co-authors V.M. Malhotra and M. Pramodh Kumar devised a max-flow algorithm that matched the best bounds at that time, but was conceptually much simpler and hence ideal for exposition.

Scheduling and facility location problems are often cast as multicommodity flow problems and are NP-hard. Using ideas from flows and linear programming, efficient approximation problems can be devised in many settings. The Indian Institute of Technology (IIT) Delhi is at the forefront of international research in this area.

Parameterized algorithms and complexity is a relatively recent field that focuses on multivariate analysis of algorithm performance and the development of algorithms for hard problems where combinatorial explosion is confined to specified parameters. This burgeoning field has a very close connection with India—the first international event wholly devoted to this theme took place in Chennai in 1999—and has seen cutting-edge contributions from India, notably from the Institute of Mathematical Sciences, Chennai (IMSc) and Chennai Mathematical Institute (CMI).

Matchings in graphs come in many different flavors—perfect, maximum, stable, popular. Indian researchers have made significant contributions toward obtaining combinatorial characterizations, devising new algorithms, and understanding the parallel complexity of these problems.

Data structures are crucial to the efficiency of many state-of-the-art algorithms. Indian researchers have been part of the community designing data structures for static succinct representations and for maintaining dynamic

IARCS
AHMEDABAD UNIVERSITY
FSTTCS 2018
38th IARCS Annual Conference on
Foundations of Software Technology and Theoretical Computer Science

The annual Foundations of Software Technology and Theoretical Computer Science (FSTTCS) Conference, organized by the Indian Association for Research in Computer Science, is a premier forum for presenting original results in initial aspects of CS and software technology. The images here show participants from FSTTCS 2018, held last December at India's Ahmedabad University.

PHOTOS COURTESY OF FOUNDATIONS OF SOFTWARE TECHNOLOGY AND THEORETICAL COMPUTER SCIENCE (FSTTCS)

data, as well as in proving non-trivial lower bounds on query complexity and space requirements.

Complexity theory. Primality testing has been studied at least since ancient Greece. However, nontrivial ideas for testing primes appeared only in the last two centuries. Apart from academic interest, primality testing has gained huge practical importance because of the need for arithmetic modulo prime and pseudo-prime numbers in various cryptographic implementations, error-correcting codes, and other fundamental computational problems.

Though randomized polynomial-time algorithms suffice for this purpose, the basic question of derandomization remained open till 2002 when the breakthrough result PRIMES is in P was proved by Agrawal et al.[1] at IIT Kanpur. Agrawal was already a well-established complexity theorist, while Kayal and Saxena were graduate students about to start their Ph.D. thesis work. This paper eventually appeared in the *Annals of Mathematics* and was awarded both the Godel Prize of EATCS-SIGACT and the Fulkerson Prize of AMS.

Algebraic complexity theory deals with the symbolic computation of formal polynomials in models such as circuits. The mathematical analysis of these models involves an interaction between computer science and algebra and enriches both fields. The recent contributions of Indian researchers at CMI, IIT Bombay, IIT Kanpur, IIT Madras, Indian Institute of Science, Bangalore (IISc), IMSc, and Microsoft Research in this technically challenging area have been stunning, with numerous foundational results and proof techniques being developed.

Algebraic methods are also used to show that certain problems are hopelessly hard by proving lower bounds. For example, the notorious problem P=NP involves proving an algorithmic lower bound. There are analogous lower bound problems for algebraic circuits. The theory research community in India has been making steady progress in this area.

Machine learning is a potential area to apply the insights gained from algebraic complexity. An artificial neural network (ANN) is an algebraic circuit with threshold gates. Thus, better understanding of threshold circuits can lead to better backpropagation algorithms and stronger lower bound results in learning theory. Indian researchers have already started designing circuit reconstruction algorithms.

Isomorphism problems about structures frequently appear in computer science. Some example structures are NP-hard problems, graphs, fields, algebras, and polynomials. Indian theorists have been studying these closely, and have proved some of the best results known.

Communication complexity studies the interaction required to solve a problem when the input is distributed across multiple parties. Indian researchers, notably at Tata Institute of Fundamental Research, Mumbai (TIFR), have made leading contributions to this area.

Logic and automata theory. The close interplay between automata theory and logic was first identified by Buchi. Pnueli introduced temporal logic as a language for specifying properties of reactive systems. Emerson, Clarke, and Sifakis invented model checking: determining algorithmically whether a formal model satisfies a temporal logic specification.

Reactive systems typically consist of many interacting components. Viewing the system as a sequential automaton results in the state explosion problem, severely limiting the effectiveness of model checking. Moreover, temporal logics interpreted over sequences are forced to reason about an exponential number of equivalent interleavings for a set of concurrent actions.

Mazurkiewicz proposed enriching alphabets with an independence relation. Adjacent independent actions commute, creating equivalence classes of words called traces. Traces are labeled partial orders of bounded width and smoothly generalize words in many respects.

Zielonka defined asynchronous automata, a distributed model that precisely captures regular trace languages. This led to a natural question of model checking asynchronous automata with respect to temporal logics defined over traces.

The first temporal logic over

In the 1980s and 1990s, theory offered a unique opportunity to keep up with international research in computing despite limited access to state-of-the-art hardware.

traces, TrPTL, was formulated in CMI. The model checking problem was solved using the gossip automaton that uses a bounded set of timestamps to dynamically keep track of updates among communicating processes.

Temporal logic is expressively equivalent to the first order theory of sequences. It is not known if TrPTL captures the first order theory of traces. Researchers at CMI, in collaboration with European colleagues, later developed the first expressively complete temporal logics over traces.

Results from trace theory generalize to communicating finite-state machines with bounded channels. Message sequence charts (MSCs) describe interactions between agents communicating through buffers. A robust theory of regular MSC languages was developed at CMI.

The converse of model checking is synthesis: construct an automaton that meets a logical specification. In the sequential setting, this was solved by Buchi and Landweber. In the distributed setting, Pnueli and Rosner proved strong undecidability results that stem from enforcing global specifications across loosely coupled agents. The decidability of distributed synthesis with local specifications is still open. Some of the strongest positive results for subclasses of systems were proved in CMI and IMSc.

Automata theory and logic have expanded to incorporate other features. A number of timed extensions to temporal logic were developed at IMSc and TIFR. In parallel, there was also work on distributed timed automata at CMI and IISc, as well as on timed versions of communicating finite-state machines at CMI and IIT Bombay. There has been work at IMSc on automata and logics over data words, which capture computations over infinite datatypes. There has also been work at CMI and IIT Kanpur in extending model checking from finite-state systems to infinite-state systems such as pushdown automata.

The Academic Ecosystem in India

Indian undergraduate programs in computing date back to the early 1980s—a time that also saw the first generation of graduate students from India taking up theoretical computer science. In the 1990s, these young researchers helped set up strong theory groups in TIFR, the IITs at Bombay, Delhi, Kanpur, and Madras, IISc, IMSc, and CMI. This network is now expanding to newer IITs at Gandhinagar, Goa, Guwahati, Hyderabad, and Palakkad, as well as IIITs and some traditional universities such as Delhi University.

The FSTTCS Conference gave rise to the Indian Association for Research in Computing Science (IARCS).[3] IARCS initiated several activities for the academic community, such as travel grants for Ph.D. students to attend conferences and faculty development programs to improve the quality of teaching. Many of these activities continue today in partnership with ACM India.

Some very robust mechanisms have arisen to sustain international collaborations. The Max-Planck Society of Germany set up the Indo-German Max Planck Center for Computer Science at IIT Delhi. The French National Centre for Scientific Research (CNRS) has established an international Research Lab in Computer Science at CMI in Chennai.

Theoretical computer science attracts some of the brightest graduate students in the country. Since the ACM India Doctoral Dissertation Awards began in 2012, nine of the 13 prizes awarded have been in theoretical computer science.

Finally, there are a large number of outstanding researchers trained in India who are active in theoretical computer science across the world. To name just two: Madhu Sudan and Subhash Khot have both won the Nevanlinna Prize awarded at the International Congress of Mathematicians.

Theoretical computer science attracts some of the brightest graduate students in the country.

References

1. Agrawal, M., Kayal,N. and Saxena, N. PRIMES is in P. *Annals of Mathematics 150* (2004), 781–793.
2. Foundations of Software Technology and Theoretical Computer Science; https://www.fsttcs.org.in/
3. Indian Association for Research in Computing Science; https://www.iarcs.org.in/

Meena Mahajan is a professor at The Institute of Mathematical Sciences, Chennai, India.

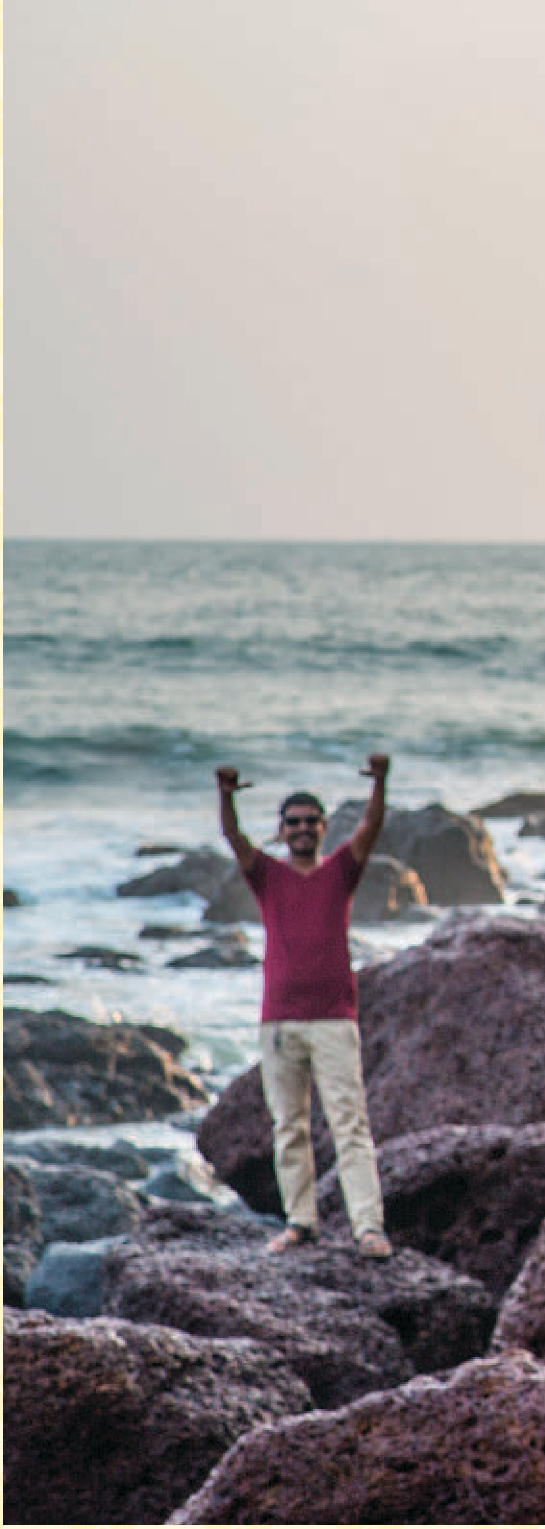
Madhavan Mukund is a professor at the Chennai Mathematical Institute, Chennai, India.

Nitin Saxena is a professor at the Indian Institute of Technology Kanpur, Kanpur, India.

DOI:10.1145/3345671

BY NILOY GANGULY AND PONNURANGAM KUMARAGURU

The Positive and Negative Effects of Social Media in India



THERE HAS BEEN a phenomenal increase in the use of online social media (OSM) services in India, including Facebook, Twitter, Instagram, LinkedIn, and YouTube. In addition to these services, one-to-one messaging services like WhatsApp have 200 million users, the highest in the world. India has 462 million users accessing the Internet, among these: Facebook has 250+ million users, LinkedIn 42+ million, and Twitter 23+ million users, and the majority of users access these services through their mobile phones.

These services have had a profound impact in India—overall digital literacy has increased, people are more connected, dissemination of local language content has increased, information exchanged during crises is substantial, and more. The deep penetration of social media services also has negative effects—the propagation of false information and hate, an increase in spammers and phishers, users are losing social skills, and more. Newness of technology/mobile phones, low-literacy rates, and cheaper mobile data rates are cited as negative impacts of social media services on society.

Research has been mainly directed toward regulation of content generated on OSM. It can be classified in the following categories:

- Identifying topical interests and expertise of the users in online behavior[1,11] and efficiently matching the consumers and producers of content;

PHOTO BY FILIP JEDRASZAK/SHUTTERSTOCK.COM

- Mining useful content from social media, for example, finding actionable information from the OSM to help law-enforcement agencies[9] and relief and rescue teams during disaster;[7,8]
- Identifying harmful content, namely analyzing hate and spam content on YouTube and Twitter,[5] and analyzing the spread of misinformation/fake content on social media (TweetCred, Facebook Inspector, WhatsFarzia[a]);
- Identifying bias in content recommendation of news to users of social media;[2]
- Impact of content on determining the dynamics of opinion over social networks.[3,4]

Note that with the rising usage of local and code-mixed (that is, local language + English) languages in content generation, a lot of research is also directed toward mining in presence of such content.[6] Selfies form a substantial part of social media image content and it has been found that clicking selfies in many cases can lead to accidents. Hence, another line of research has been to accurately communicate to users risks involved with a location chosen for taking selfies, as with the Saftie and Saftie Camera apps.[b]

Research enumerated earlier provides an overview of some of the ongoing work in the area of social media conducted by Indian scientists, but is by no means exhaustive. Here,

[a] http://precog.iiitd.edu.in/research/whatsapp-misinformation/

[b] http://labs.precog.iiitd.edu.in/killfie/

we elaborate on some of the work, specifically focusing on a set of work that helps users get access to 'useful' and 'sanitized' content. We will also talk about the issues related to code-mixed text and the specific research undertaken to identify dangerous spots for clicking selfies.

A lot of research is directed toward code-mixed content, which combines a local language and English.

Search and recommendation systems over OSM. In order to develop search and recommendation systems over OSMs, it is critical to have accurate methodologies for tasks like inferring the topical interests and expertise of users, and searching for experts on specific topics. Researchers proposed completely novel crowdsourcing-based methodologies for these tasks, for example, the topics of expertise of a user are inferred based on how other users describe the said user.

The proposed methodologies are far more accurate than content-based techniques, in inferring a wide range of topics of interest/expertise of users and identifying topical experts. It was earlier thought that OSMs like Twitter are only used for casual conversation among friends. However, several works[1,11] showed that Twitter is actually a treasure-trove of information on thousands of topics, ranging from popular topics like politics and sports, to specialized topics like neurology and forensics. The research has identified thousands of groups of Twitter users interested in these diverse topics. Along with proposing novel algorithms, the endeavor has resulted in the development and public deployment of several Web-based systems on the Twitter platform based upon the proposed algorithms, for example, topical search systems,[c] systems for inferring topical interest/expertise of users,[d] and so on. These systems are currently being used by hundreds of users worldwide.

Efficient utilization of social media during disasters. Research has shown that microblogging sites like Twitter have become important sources of real-time information during disaster events. A significant amount of valuable *situational information* (updates about a current situation) is available from these sites. However, this information is immersed among hundreds of thousands of tweets, mostly containing sentiments and opinions of the masses who are posting during such events. To effectively utilize microblogging sites during disaster events, a series of research work conducted by CNeRG IIT Kharagpur[e] has extracted the situational information from among the large amounts of sentiment and opinion, determined the humanitarian categories like 'infrastructure damage,' 'missing or found people,' or 'relief required' of the tweets, and summarized the situational information in real time, to help decision-making processes when time is critical.

Another important observation is that apart from English, people also post situational updates in their local languages (predominantly Hindi in India)—hence the classification-summarization framework was extended to Hindi as well as code-mix (for example, part Hindi, part English) tweets. It has also been observed that some people take advantage of a panic situation, posting offensive content targeting specific religious communities during a disaster. Such communal posts deteriorate law and order and unfortunately it has been observed on the Indian subcontinent that this phenomenon is prevalent even during a natural disaster. Methods to detect such communal tweets and to characterize users who initiate and/or propagate them were developed.

Election and social media: Researchers in India have studied in detail the use of social media during the April/May 2019 elections in India and made several observations.[f] Besides the widespread usage of misleading messages and suspected (fake/bot) accounts, which are now observed in almost all elections, there were several specialties, including a substantial amount of satire video; female verified handles demonstrate more engagement compared to male verified accounts; and an important trending hashtags has been #MainBhiChowkidar (#IamtheWatchMan), which prompted around 5,000 users to add Chowkidar (Watchman) to their name in the social media handle.

Code mixing on social media. There

[c] http://bit.ly/2kf9NGy and http://bit.ly/2lWeYMk
[d] http://bit.ly/2kCIZ3u and http://bit.ly/2kOJRSm
[e] http://www.cnerg.org
[f] http://labs.precog.iiitd.edu.in/elections-2019

is a widespread practice of writing Indian languages using Roman script as well as mixing it with English during writing/speaking,[g] a phenomenon referred to as *linguistic code-mixing* or *code-switching*. For any analysis of social media content from India, correct processing of code-mixed text is an absolute necessity; however, traditional natural language processing (NLP) modules such as language identifiers, POS taggers, translators, and word aligners treat linguistic code-switching data either as noise or as a new language (for example, Hinglish for Hindi-English code mixing). Both views are limited because the former does not recognize the complexity and socio-pragmatics of the phenomenon, whereas the latter does not utilize the fact that code mixing is a grammatically informed combination of two languages. Further, bilingual speakers show different language references depending on the topic of discussion and sentiment expressed. This implies that ignoring code-mixed patterns or conducting content-analysis only for the predominant language over social media (usually English) can lead to misleading conclusions, and are bound to miss out on social and discourse-level nuances in the data. Several researchers from India have worked to address different aspects of code-switching; Microsoft Research India, under project Melange,[h] has largely led the initiative. Several semi-supervised[10] techniques to automatically produce a large, annotated code-mixed dataset are being developed to help the community efficiently perform downstream supervised NLP tasks.

Killfies for social media. In recent years, the posting of selfies (or digital self-portraits) on social media websites such as Facebook, Instagram, and Snapchat has become a part of mainstream culture. Often people portray their adventurousness by posting dangerous selfies (aka killfies). Since March 2014, 238 people are reported to have been killed while taking selfies,[i] with India dominating these statistics with 141 deaths. Given the increasing penetration of mobile technology, high usage statistics, and the disturbances caused by such behavior, India is one of the prime regions where this problem is particularly relevant. Research conducted by Precog@IIIT Delhi[j] identifies dangerous selfies. The researchers have created datasets, classifiers, apps, and location-marker tools in this context. A convolutional neural network-based classifier to identify dangerous selfies posted on social media using only the image (no metadata) gives an accuracy of 98%. The Saftie Camera[k] app based on the developed classifier works in real world settings and detects and warns a user if the location is potentially dangerous.

Important funding initiatives. There has been a lot of funding initiatives both from government and non-government agencies to popularize social media research. Among those initiatives is the Indo-German Max Planck Center for Computer Science—a five-year project on *Understanding, leveraging and deploying online social networks*, jointly funded by the Indian Department of Science and Technology and Max Planck Society. Another initiative is the Media Lab Asia and Information Technology Research Academy (ITRA)-funded five-year project on *Post disaster situation analysis and resource management*, which patronized the research on investigating the role of social media for disaster management.

Challenges. Presently, the world is witnessing several negative impacts of OSMs. Hence, it is important for the computing world, with intense research input from scientists all over the world, to mitigate these impacts. The specific problems are many—fake news, hate speech, the shaming of individuals or groups. It is now clear that in the garb of spontaneity, companies, political parties, and individuals are constantly manipulating the systems to produce trending topics and thus control discussions on social media. The problems are compounded in India with the unprecedented rise in use of local or code-mix languages; hence the need for special attention from Indian researchers. Another diagonally opposite area of research would be to leverage social media for social good; work on post-disaster management as reported here; and future scopes including utilizing social media content to devise better governance mechanisms, supporting individuals/groups with health-related issues, and making quality education accessible to the huge population by connecting teachers with students located in different places.

Acknowledgments. The authors thank Sunita Sarawagi, Abir De, and the anonymous reviewers for providing constructive feedback.

References

1. Bhattacharya, P. et al. Deep Twitter diving: Exploring topical groups in microblogs at scale. In *Proceedings of the 17th ACM Conf. Computer Supported Cooperative Work and Social Computing*, 2014, 197–210.
2. Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N.and Gummadi, K.P. Who makes trends? Understanding demographic biases in crowdsourced recommendations. In *Proceedings of the 11th Intern. AAAI Conf. Web and Social Media*, 2017.
3. De, A., Bhattacharya, S.and Ganguly, N. Demarcating Endogenous and Exogenous Opinion Diffusion Process on Social Networks. In *Proceedings of the 2018 World Wide Web Conf.*, 2018, 549–558.
4. De, A., Valera, I., Ganguly, N., Bhattacharya, S. and Gomez-Rodriguez, M. Learning and forecasting opinion dynamics in social networks. In *Proceedings of the 30th Inter. Conf. Neural Information Processing Systems*, 2016, 397–405.
5. Maity, S.K., Chakraborty, A., Goyal, P. and Mukherjee, A. Opinion conflicts: An effective route to detect incivility in Twitter. In *Proc. ACM Hum.-Comput. Interact. Article 117* (2018), 117:1–117:27.
6. Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S. and Bali, K. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Assoc. Computational Linguistics, Vol.1.* (Melbourne, Australia, 2018), 1543–1553; https://www.aclweb.org/anthology/P18- 1143
7. Rudra, K., Ganguly, N., Goyal, P. and Ghosh, S. Extracting and summarizing situational information from Twitter social media during disasters. *ACM Trans. Web 12*, 3 (July 2018), 17:1–17:35.
8. Rudra, K., Goyal, P., Ganguly, N., Mitra, P. and Imran, M. Identifying sub-events and summarizing disaster-related information from microblogs. In *Proceedings of the 41st Intern. ACM SIGIR Conf. Research and Development in Info. Retrieval*, 2018, 265–274.
9. Sachdeva, N. and Kumaraguru, P. Call for service: Characterizing and modeling police response to serviceable requests on Facebook. In *Proceedings of the ACM Conf. Computer-Supported Cooperative Work and Social Computing*, 2017.
10. Samanta, N., Nangi, S.R., Jagirdar, H., Ganguly, N., Charabarti, S. A deep generative model for code switched text. In *Proceedings of IJCAI*, 2019.
11. Zafar, M.B., Bhattacharya, P., Ganguly, N., Ghosh, S. and Gummadi, K.P. On the wisdom of experts vs. crowds: Discovering trustworthy topical news in microblogs. In *Proceedings of the ACM Conf. Computer-Supported Cooperative Work and Social Computing*, 2016, 438–451.

Niloy Ganguly (niloy@cse.iitkgp.ac.in) is a professor at IIT Kharagpur, India.

Ponnurangam Kumaraguru (pk@iiitd.ac.in) is an associate professor at IIIT Delhi, India.

[g] for example, a bilingual Hindi/English speaker posts on Twitter: "*aj patakhe to india me hi phutenge*, sure it would be," where the *italicized segment* ("today fireworks will occur in India only") is in Hindi written in Roman script.

[h] https://www.microsoft.com/en-us/research/project/melange/

[i] http://bit.ly/saftie- bot

[j] http://precog.iiitd.edu.in/

[k] http://bit.ly/saftie-cam

DOI:10.1145/3343452

BY ADITYA VASHISTHA, UMAR SAIF, AND AGHA ALI RAZA

The Internet of the Orals

INTERNET SERVICES LIKE social media, online discussion forums, and crowdsourcing marketplaces have transformed how people participate in the information ecology and digital economy. These services empower mostly urban, affluent, and literate people, and improve their reach to information and instrumental needs. However, these services currently exclude billions of people worldwide who are too poor to afford Internet-enabled devices, too remote to access the Internet, or too low literate to navigate the mostly text-driven Internet.

In India and Pakistan alone, there are nearly 1.1 billion people offline. Although 70% of their populations have access to mobile phones, most people still use basic or feature phones, making it difficult to extend existing Internet services on these devices running custom operating systems. Even when people can afford smartphones and the Internet, literacy barriers prevent 26% of adults in India and 42% of adults in Pakistan from using text-based interfaces. Most South Asian languages and dialects are still unsupported by the advancements in natural language processing ruling out the use of voice interfaces like Siri and Alexa.

In light of these constraints, Human-Computer Interaction for Development (HCI4D) researchers and practitioners have used interactive voice response (IVR) technology to create voice-based services that overcome connectivity barriers by using ordinary phone calls, literacy barriers by using local language speaking and listening skills, and socioeconomic barriers by using toll-free (1-800) lines. These services let users call a phone number to record and listen to voice messages in their local languages. Because of their accessible and usable design, these services have found applications in diverse domains and have profoundly impacted marginalized communities in low-resource environments. This article follows the evolution of these services over the last two decades (see the accompanying figure), and their big challenges and new frontiers.

First Wave: Access and Inclusion

The first wave of voice-based services focused on improving information access for people in low-resource communities. For example, HealthLine enabled low-literate frontline health workers in Pakistan to retrieve relevant information by speaking out predefined commands.[6] While initial efforts like HealthLine allowed users to only consume information, subsequent services took the form of voice forums and enabled marginalized communities to also produce and share information. This included Avaaj Otalo (an agriculture discussion forum in India),[3] CGNet Swara (a citizen journalism service in India),[2] MobileVaani (a social media service in India), Ila Dhageyso (a civic engagement portal in Somaliland),[1] and IBM's Spoken Web (a user-generated

A blind user of Sangeet Swara recording a voice message.

information directory in India). The success of these initial services demonstrated their great potential to enable information access and connectivity among underserved populations in diverse HCI4D contexts. However, the vast majority of these services ran into the hurdles of user training and technology adoption.

Second Wave: Training and Spread

Nearly a decade ago, the biggest roadblocks to designing voice forums were usability, motivation, and spread; target populations faced difficulties in using even the simplest of speech-based telephone interfaces, they did not exhibit interest or trust in using such services, and it was difficult to advertise and spread such services to underconnected people. Researchers tried to overcome these barriers by conducting lab-trainings as well as door-to-door field campaigns, but it was quickly realized that these approaches were not scalable. Raza et al. used a ludic design approach to train users and promote usability and spread. They built Polly, a voice-based entertainment service that lets users make a short audio recording, apply funny voice modifications to it, and share it with their friends via automated voice calls.[5] They deployed Polly to five low-income people in Pakistan in early 2012. Within a year, Polly spread virally to over 165,000 users via 636,000 calls without any outreach efforts. Polly's ludic interfact design trained users to navigate IVR interfaces, and also led to its viral adoption. Raza et al. then used Polly to share instrumental information with users to aid their socioeconomic development. In an initial test, 34,000 Polly users listened to 728 job advertisements nearly 386,000 times within a year.

Over the last seven years, Polly has been successfully used in multiple countries to rapidly spread useful information to underserved populations. In 2014, at the peak of the Ebola crisis in West Africa, Polly-Santé (Polly-Health) was deployed as an emergency disaster-response service in Guinea to spread reliable information about prevention, symptoms, and cure of Ebola.[12] The information originated from the Centers for Disease Control and the service was funded by the U.S. Embassy in Conakry. One of the hurdles to information dissemination in the Guinean context is great linguistic diversity and the lack of a widely understood

Because of their accessible and usable design, voice-based services have found applications in diverse domains and have profoundly impacted marginalized communities in low-resource environments.

common language. Fortunately, this is not a major impediment for voice forums. Polly-Santé was launched in 11 local languages and reached more than 7,000 local mobile phone users within a few months. In 2014, Polly was also used in India by Babajob.com to advertise a voice directory of available jobs to thousands of low-literate job seekers.

Since 2016, Polly has been active in Pakistan as a gateway to maternal health information for underconnected expectant parents. Polly advertises a hotline called Super Abbu (Super Dad) that allows expectant parents to record health questions that are answered by volunteer doctors. Such private and anonymous access to trained gynecologists allows parents to ask questions about pregnancy and childbirth that are often considered sensitive and even taboo topics in the local context. The service specifically targets fathers to promote paternal participation and allow them to share their experiences with their peers. In its initial deployment, Super Abbu reached 21,000 users (96% of them men) in just two months, uncovering a pent-up demand for maternal health information and giving the target population an agency to anonymously access culturally sensitive yet lifesaving reliable information.

Despite their demonstrated impact, large-scale voice forums like Polly face two challenges that significantly impede their scalability and sustainability: how to manage user-generated content in local languages, and how to manage the cost of voice calls from users to access these services.

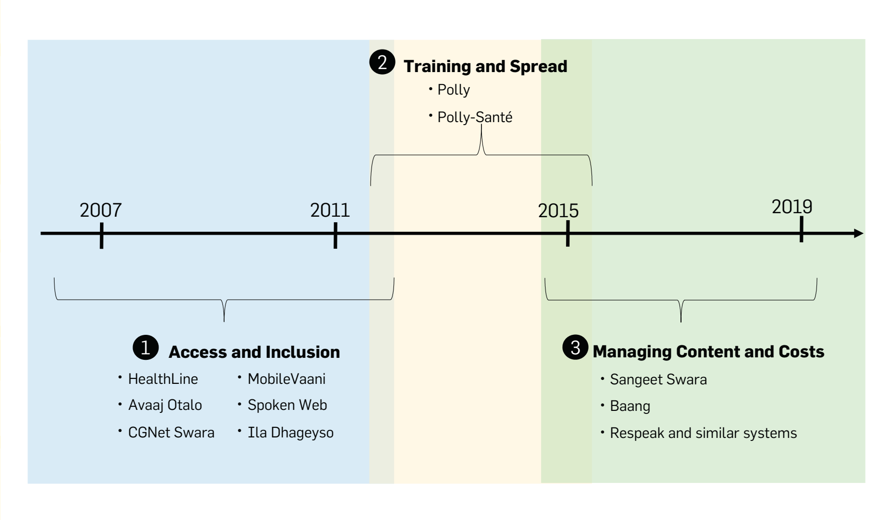
Third Wave: Managing Content and Costs at Scale

Voice forums deployed in low-resource environments often receive large volumes of user-recorded content in local languages and accents that have no speech corpora and recognition models. Consequently, it is very difficult to moderate, search, and index such content at large scale. Various voice forums often hire a dedicated team of moderators who listen to messages, categorize them, and review the quality. However, manual moderation is difficult to scale if these services grow by orders of magnitude. To address this challenge, Vashistha et al. harnessed crowdsourcing and showed that the users of voice forums, although socioeconomically marginalized and technologically inexperienced, can themselves be entrusted with the tasks of audio content moderation and categorization. In 2014, they built Sangeet Swara, a community-moderated social media voice forum that lets users record, listen to, and vote on songs, poems, and other cultural content.[10] As users listen to messages, Swara requests them to annotate the quality and category by pressing phone keys (for example, press 1 to upvote or 2 to downvote the message) and uses collaborative filtering techniques to rank, order, and categorize audio messages based on users' votes.

In an eight-month deployment in India, Swara received 53,000 phone calls from 13,000 users who submitted 6,000 voice messages in 11 languages as well as 150,000 votes. Nearly 80% of users had never used any social media platform before, 50% lived in low-income environments in rural India, and 25% were people with vision impairments (as shown in the opening image). Community moderation was 98% accurate in content categorization, made meaningful distinctions between high- and low-quality posts, and performed judgments that were in 90% agreement with expert moderators.

Deriving inspiration from Swara, Raza et al. used community moderation to manage content on Baang, a voice-based social media platform that encouraged users to record and share audio messages of diverse genres.[4] Baang allowed users to also record threaded audio comments on voice messages and added a Polly-like sharing mechanism. Deployed in Pakistan in 2015, Baang organically reached 10,000 users within eight months who contributed more than 44,000 voice messages that were played more than 2.8 million times, and received nearly 340,000 votes and 124,000 audio comments. The ability to vote, comment, and share led to viral spread, deeper engagement, and the emergence of true dialog among participants. Beyond connectivity, Swara and Baang provided its users with a voice and a social identity as well as a means to share informa-

Three waves of voice forums in low-resource environments.



tion and get community support. Moreover, they demonstrated that a community of low-income, low-literate people can moderate themselves without any outside support, thereby addressing the content management challenge of these voice forums.

The second key challenge in scaling voice forums is the airtime cost. Often, these services use expensive toll-free lines to remain accessible to low-income users. The resultant cost poses a huge burden to sustainability, often putting these services at risk of being shut down as the usage grows. While a few services sustain themselves through advertisements, grants, and partnerships with telecoms or governments, these options are often beyond the reach of most voice forum providers. To make these services financially sustainable, Vashistha et al. examined whether low-income users of voice forums could complete useful work on their mobile phones to offset their participation costs. In 2016, they created Respeak, the first voice-based crowdsourcing marketplace that pays users to transcribe audio files vocally.[7–9] Respeak sends short audio segments to multiple voice forum users and pays them via mobile airtime for each submitted transcript. Instead of typing the transcript, users respeak audio content into an off-the-shelf speech recognition engine and submit the autogenerated transcript. Respeak combines the transcripts for each segment from multiple users using sequence-alignment algorithms to reduce random speech recognition errors. It then pays users in mobile airtime based on the accuracy of transcripts submitted in them. In the last three years, Respeak has been used by low-income students, blind people, and rural residents in India to produce speech transcriptions with over 90% accuracy at one-fourth of the market rate, generating sufficient profit to subsidize their participation costs. One minute of crowd work on Respeak enable users to earn eight minutes of airtime.[8]

Grand Challenges: Harassment, Misinformation, and Disinformation

Voice forums, like any other social platform, come with their own pitfalls. They end up reflecting the existing sociocultural norms and values of the society, including its shortcomings and biases. For example, while Swara and Baang served as instruments of inclusion for low literate, rural, indigenous, and visually impaired communities, they failed to create a welcoming environment for female users.[11] Women faced systemic discrimination and harassment in the form of messages that contained abuses, threats, and flirtatious behavior.

Both mainstream social media platforms and voice forums face grand challenges when tackling misinformation, disinformation, harassment, and abuse. These platforms and forums differ greatly in terms of scale, features, interfaces, supported languages, and target users. Consequently, solutions to tackle these challenges on a platform like Facebook might be ineffective for voice forums, and vice versa. This presents interesting research challenges of identifying indecorous content in local language audio, filtering out spreaders of disinformation, and addressing situations where the collective ignorance of community members eclipse their collective intelligence. The HCI4D community must tackle these grand challenges to make the Internet of the orals more diverse, inclusive, and impactful.

References

1. Gulaid, M. and Vashistha, A. Ila Dhageyso: An interactive voice forum to foster transparent governance in Somaliland. In *Proceedings of the 6th Intern. Conf. Information and Communications Technologies and Development: Notes, Vol. 2* (Cape Town, South Africa, 2013), 41–44.
2. Mudliar, P. et al. Emergent practices around CGNet Swara, voice forum for citizen journalism in rural India. In *Proceedings of the 5th Intern. Conf. Information and Communication Technologies and Development* (Atlanta, GA, USA, 2012), 159–168.
3. Patel, N. et al. Avaaj Otalo: A field study of an interactive voice forum for small farmers in rural India. In *Proceedings of the SIGCHI Conf. Human Factors in Computing Systems* (Atlanta, GA, USA, 2010), 733–742.
4. Raza, A.A. et al. Baang: A viral speech-based social platform for under-connected populations. In Proceedings of the 2018 CHI Conf. Human Factors in Computing Systems (Montreal, QC, Canada, 2018), 643:1–643:12.
5. Raza, A.A. et al. Job opportunities through entertainment: Virally spread speech-based services for low-literate users. In *Proceedings of the SIGCHI Conf. Human Factors in Computing Systems* (Paris, France, 2013), 2803–2812.
6. Sherwani, J. et al. Healthline: Speech-based access to health information by low-literate users. *Inter. Conf. Information and Communication Technologies and Development* (Bangalore, India, 2007), 1–9.
7. Vashistha, A. et al. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conf. Human Factors in Computing Systems* (Montreal, QC, Canada, 2018), 57:1–57:13.
8. Vashistha, A. et al. ReCall: Crowdsourcing on basic phones to financially sustain voice forums. In *Proceedings of the 2019 CHI Conf. Human Factors in Computing Systems* (Glasgow, Scotland, U.K., 2019).
9. Vashistha, A. et al. Respeak: A voice-based, crowd-powered speech transcription system. In *Proceedings of the 2017 CHI Conf. Human Factors in Computing Systems* (Denver, CO, USA, 2017), 1855–1866.
10. Vashistha, A. et al. Sangeet Swara: A community-moderated voice forum in rural India. In *Proceedings of the 33rd Annual ACM Conf. Human Factors in Computing Systems* (Seoul, South Korea, 2015), 417–426.
11. Vashistha, A. et al. Threats, abuses, flirting, and blackmail: Gender inequity in social media voice forums. In *Proceedings of the 2019 CHI Conf. Human Factors in Computing Systems* (Glasgow, Scotland, U.K., 2019).
12. Wolfe, N. et al. Rapid development of public health education systems in low-literacy multilingual environments: Combating Ebola through voice messaging. In *Proceedings of the ISCA Special Interest Group on Speech and Language Technology in Education* (Leipzig, Germany, 2015).

Aditya Vashistha is an assistant professor at Cornell University, Ithaca, NY, USA.

Umar Saif is UNESCO Chair, ICTD, Lahore, Pakistan.

Agha Ali Raza is an assistant professor at Information Technology University, Lahore, Pakistan.

DOI:10.1145/3359338

Article development led by acmqueue
queue.acm.org

Collaboration between humans and machines does not necessarily lead to better outcomes.

BY MICHELLE VACCARO AND JIM WALDO

The Effects of Mixing Machine Learning and Human Judgment

IN 1997, IBM'S Deep Blue software beat the World Chess Champion Garry Kasparov in a series of six matches. Since then, other programs have beaten human players in games ranging from "Jeopardy!" to Go. Inspired by his loss, Kasparov decided in 2005 to test the success of Human+AI pairs in an online chess tournament.[2] He found the Human+AI team bested the solo human. More surprisingly, he also found the Human+AI team bested the solo computer, even though the machine outperformed humans.

Researchers explain this phenomenon by emphasizing that humans and machines excel in different dimensions of intelligence.[9] Human chess players do well with long-term chess strategies, but they perform poorly at assessing the millions of possible configurations of pieces. The opposite holds for machines. Because of these differences, combining human and machine intelligence produces better outcomes than when each works separately. People also view this form of collaboration between humans and machines as a possible way to mitigate the problems of bias in machine learning, a problem that has taken center stage in recent months.[12]

We decided to investigate this type of collaboration between humans and machines using risk-assessment algorithms as a case study. In particular, we looked at the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, a well-known (perhaps infamous) risk-prediction system, and its effect on human decisions about risk. Many state courts use algorithms such as COMPAS to predict defendants' risk of recidivism, and these results inform bail, sentencing, and parole decisions.

Prior work on risk-assessment algorithms has focused on their accuracy and fairness, but it has not addressed their interactions with human decision makers who serve as the final arbitrators. In one study from 2018, Julia Dressel and Hany Farid compared risk assessments from the COMPAS software and Amazon Mechanical Turk workers, and found that the algorithm and the humans achieved similar levels of accuracy and fairness.[6] This study signals an important shift in the literature on risk-assessment instruments by incorporating human subjects to contextualize the accuracy and fairness of the algorithms. Dressel and Farid's study, however, divorces the human decision makers and the algorithm when, in fact, the current model indicates that humans and algorithms would work in tandem.

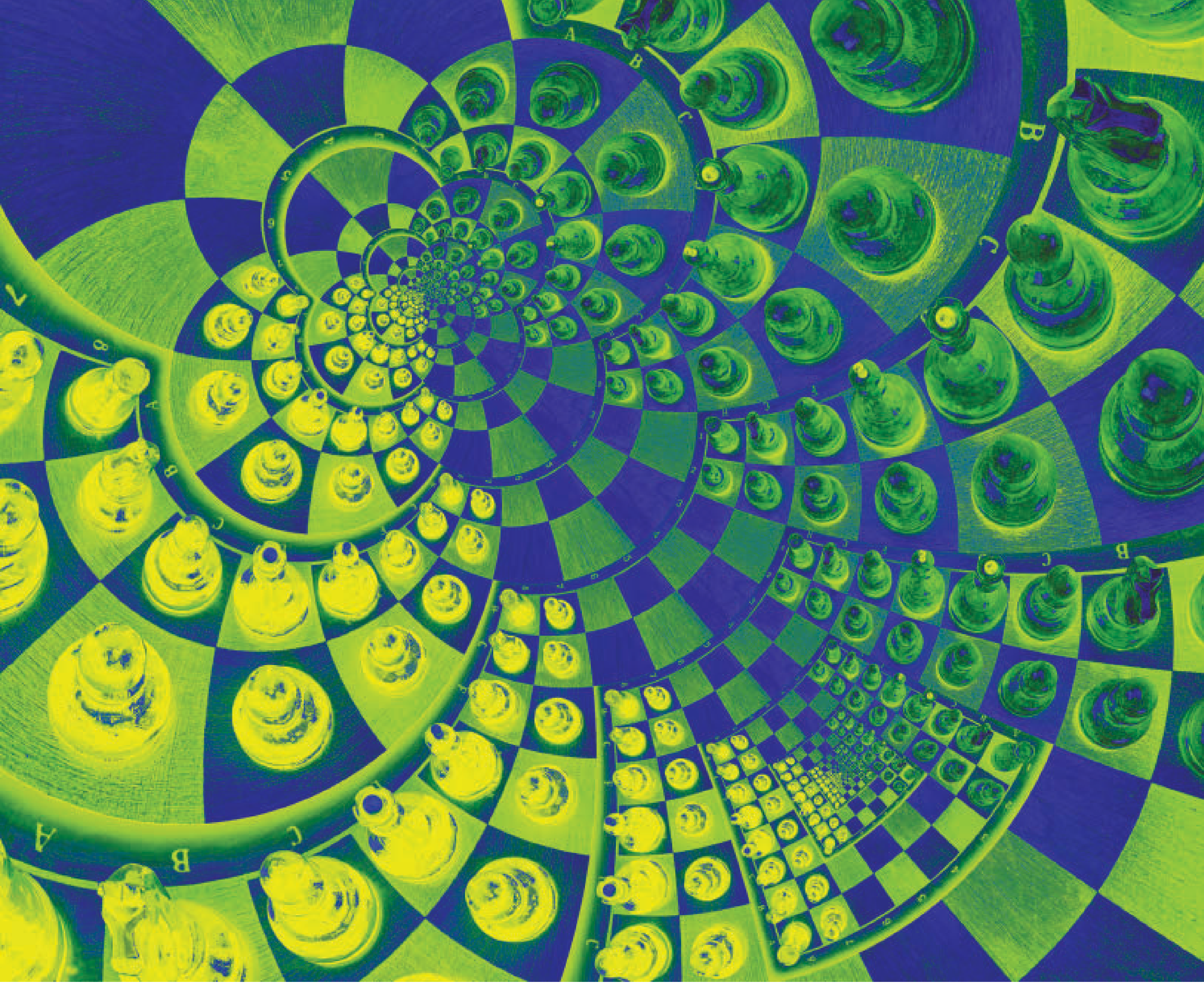


IMAGE BY MIKHAIL LEONOV, COLORIZED BY ANDRIJ BORYS ASSOCIATES

Our work, consisting of two experiments, therefore first explores the influence of algorithmic risk assessments on human decision making and finds that providing the algorithm's predictions does not significantly affect human assessments of recidivism. The follow-up experiment, however, demonstrates that algorithmic risk scores act as anchors that induce a cognitive bias: If we change the risk prediction made by the algorithm, participants assimilate their predictions to the algorithm's score.

The results highlight potential shortcomings with the existing human-in-the-loop frameworks. On the one hand, when algorithms and humans make sufficiently similar decisions their collaboration does not achieve improved outcomes. On the other hand, when algorithms fail, humans may not be able to compensate for their errors. Even if algorithms do not officially make decisions, they anchor human decisions in serious ways.

Experiment One: Human-Algorithm Similarity, not Complementarity

The first experiment examines the impact of the COMPAS algorithm on human judgments concerning the risk of recidivism. COMPAS risk scores were used because of the data available on that system, its widespread usage in prior work about algorithmic fairness, and the use of the system in numerous states.

Methods. The experiment entailed a 1 x 3 between-subjects design with the following treatments: *control*, in which participants see only the defendant profiles; *score*, in which participants see the defendant profiles and the defendant COMPAS scores; and *disclaimer*, in which participants see the defendant profiles, the defendant COMPAS scores, and a written advisement about the COMPAS algorithm.

Participants evaluated a sequence of defendant profiles that included data on gender, race, age, criminal charge, and criminal history. These profiles described real people arrested in Broward County, FL, based on information from the dataset that *ProPublica* used in its analysis of risk-assessment algorithms.[1] While this dataset originally contained 7,214 entries, this study applied the following filters before sampling for 40 profiles that were presented to participants:

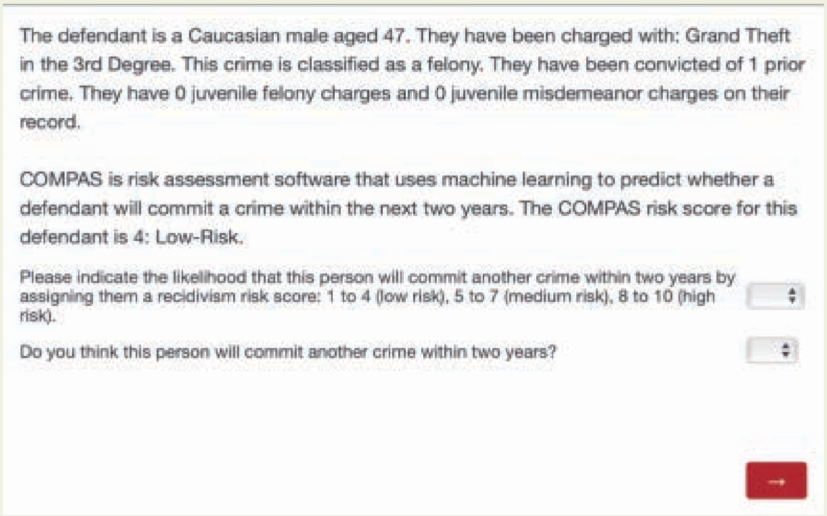
▸ *Limit to black and white defendants.* Prior work on the accuracy and fairness of the COMPAS algorithm limits their analyses to white and black defendants.[3,4,6] To compare the results from this experiment with those in prior studies, this study considers only the subset of defendants who identify as either African-American (black) or Caucasian (white).

▸ *Exclude cannabis crimes.* Interestingly, the pilot study showed participant confusion about cannabis-related crimes such as possession, purchase, and delivery. In the free-response section of the survey, participants made comments such as "Cannabis is fully legal here." To avoid confusion about the legality of cannabis in various states, this study excludes defendants charged with crimes containing the term *cannabis*.

From this filtered dataset 40 defendants were randomly sampled. A profile was generated containing information about the demographics, alleged crime, criminal history, and algorithmic risk assessment for each of the defendants in the sample. The descriptive paragraph in the control treatment assumed the following format, which built upon that used in Dressel and Farid's study:[6]

The defendant is a [RACE] [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE-FELONY COUNT] juvenile felony charges and [JUVENILE-MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.

Figure 1. Defendant profile from score treatment.



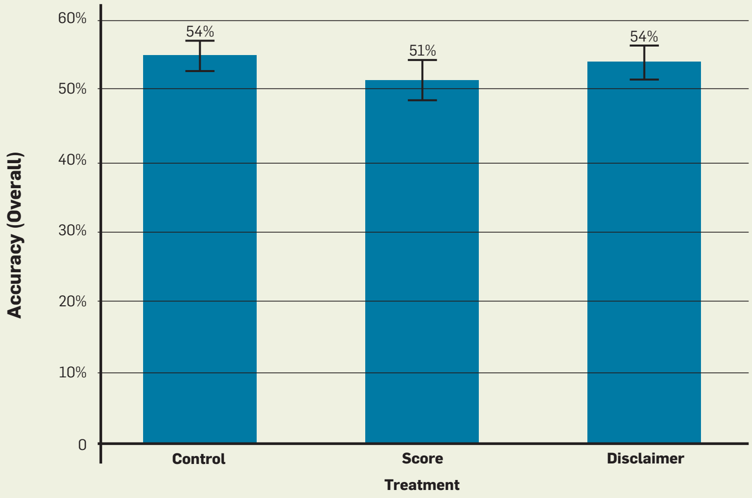
The defendant is a Caucasian male aged 47. They have been charged with: Grand Theft in the 3rd Degree. This crime is classified as a felony. They have been convicted of 1 prior crime. They have 0 juvenile felony charges and 0 juvenile misdemeanor charges on their record.

COMPAS is risk assessment software that uses machine learning to predict whether a defendant will commit a crime within the next two years. The COMPAS risk score for this defendant is 4: Low-Risk.

Please indicate the likelihood that this person will commit another crime within two years by assigning them a recidivism risk score: 1 to 4 (low risk), 5 to 7 (medium risk), 8 to 10 (high risk).

Do you think this person will commit another crime within two years?

Figure 2. Accuracy rate in treatment groups.



The descriptive paragraph in the score treatment added the following information:

COMPAS is risk-assessment software that uses machine learning to predict whether a defendant will commit a crime within the next two years. The COMPAS risk score for this defendant is [SCORE NUMBER]: [SCORE LEVEL].

Finally, the descriptive paragraph in the disclaimer treatment provided the following information below the COMPAS score, which mirrored the language the Wisconsin Supreme Court recommended in *State v Loomis*:[18]

Some studies of COMPAS risk-assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.

Upon seeing each profile, participants were asked to provide their own risk-assessment scores for the defendant and indicate if they believed the defendant would commit another crime within two years. Using dropdown menus, they answered the questions shown in Figure 1.

We deployed the task remotely through the Qualtrics platform and recruited 225 respondents through Amazon Mechanical Turk, 75 for each treatment group. All workers could view the task title, "Predicting Crime;" task description, "Answer a survey about predicting crime;" and the key words associated with the task, "survey, research, and criminal justice." Only workers living in the U.S. could complete the task, and they could do so only once. During the pilot study among an initial test group of five individuals, the survey required an average of 15 minutes to complete. As the length and content of the survey resembled that of Dressel and Farid's,[6] we adopted their payment scheme, giving workers $1 for completing the task and a $2 bonus if the overall accuracy of the respondent's predictions exceeded 65%. This payment structure motivated participants to pay close attention and provide their best responses throughout the task.[6,17]

Results. Figure 2 shows the average accuracy of participants in the control, score, and disclaimer treatments.

The error bars represent the 95% confidence intervals. The results suggest the provision of COMPAS scores did not significantly affect the overall accuracy of human predictions of recidivism. In this experiment, the overall accuracy of predictions in the control treatment (54.2%) did not significantly vary from those in the score treatment (51.0%) ($p = 0.1460$).

The inclusion of a written advisement about the limitations of the COMPAS algorithm did not significantly affect the accuracy of human predictions of recidivism, either. Participants in the disclaimer treatment achieved an average overall accuracy rate of 53.5%, whereas those in the score condition achieved 51.0%; a two-sided *t*-test indicated this difference was not statistically significant ($p = 0.1492$).

Upon the conclusion of the task block in the exit survey, 99% of participants responded that they found the instructions for the task clear, and 99% found the task satisfying. In their feedback, participants indicated they had positive experiences with the study, leaving comments such as: "I thoroughly enjoyed this task;" "It was a good length and good payment;" and "Very good task."

Participants did not mention the advisement when asked how they took the COMPAS scores into account. Rather, their responses demonstrated that they used the COMPAS scores in different ways: some ignored them, some relied heavily on them, some used them as starting points, and others used them as sources of validation.

Figure 3 has excerpts of participant responses with a summary of answers to the free-response question: How did you incorporate the COMPAS risk scores into your decisions?

Discussion. When assessing the risk that a defendant will recidivate, the COMPAS algorithm achieves a significantly higher accuracy rate than participants who assess defendant profiles (65.0% vs. 54.2%). The results from this experiment, however, suggest that merely providing humans with algorithms that outperform them in terms of accuracy does not necessarily lead to better outcomes. When participants incorporated the algorithm's risk score into their decision-making process, the accuracy rate of their predictions did not significantly change. The inclusion of a written advisement providing information about potential biases in the algorithm did not affect participant accuracy, either.

Given research in complementary computing that shows coupling human and machine intelligence improves their performance,[2,9,11] this finding seems counterintuitive. Yet successful instances of human and machine collaboration occur under circumstances in which humans and machines display different strengths. Dressel and Farid's study demonstrates the striking similarity between recidivism predictions by Mechanical Turk workers and the COMPAS algorithm.[6] This similarity may preclude the possibility of complementarity. Our study reinforces this similarity, indicating the combination of human and algorithm is slightly (although not statistically significantly) worse than the algorithm alone and similar to the human alone.

Moreover, this study shows that the accuracy of participant predictions of recidivism does not significantly change when a written advisement about the appropriate usages of the COMPAS algorithm is included. The Wisconsin Supreme Court mandated the inclusion of an advisement without indicating that its effect on officials' decision-making was tested.[11] Psychology research and survey-design literature indicate that people often skim over such disclaimers, so they do not perform their intended purpose.[10] In concurrence with such theories, the results here suggest that written advisements accompanying algorithmic outputs may not affect the accuracy of decisions in a significant way.

Experiment Two: Algorithms as Anchors

The first experiment suggested that COMPAS risk scores do not impact human risk assessments, but research in psychology implies that algorithmic

Figure 3. Participant responses to free-response question.

	COMPAS	Disclaimer
Ignore	"I tried *not* to look at them after awhile, because I felt some were off (lol) but I still took them into account somewhat. I mostly went with my gut and opinions, though."	"I thought it was fairly random, so I didn't invest much faith in it." "Generally I just ignored it and made my own guess."
Rely Heavily	"I kept my scores within 2 points of the COMPAS score."	"I relied on it, it eliminates bias."
Starting Point	"I used that as a baseline."	"It was only a starting point. I paid more attention to the criminal charge, prior charges and guessing on whether the defendant would be convicted." "I took a look at the risk score to ballpark it."
Validation	"I used to judge my final answer." "I compared my answer to their answer and adjusted mine slightly based on theirs."	"I made my own guess on what I thought and then checked the COMPAS to find a score I was happy with." "I used the COMPAS core to verify my decision."
Factor	"I used it in combination with other factors to make my ultimate decision." "I used them in consideration with the other data provided."	"I tended to look at them but used the seriousness of crime or amount of past crimes and age as my main deciding factor." "I took it into consideration along with all the other information given."

predictions may influence humans' decisions through a subtle cognitive bias known as the *anchoring effect*: when individuals assimilate their estimates to a previously considered standard. Amos Tversky and Daniel Kahneman first theorized the anchoring heuristic in 1974 in a comprehensive paper that explains the psychological basis of the anchoring effect and provides evidence of the phenomenon through numerous experiments.[19] In one experiment, for example, participants spun a roulette wheel that was predetermined to stop at either 10 (low anchor) or 65 (high anchor). After spinning the wheel, participants estimated the percentage of African nations in the United Nations. Tversky and Kahneman found that participants who spun a 10 provided an average guess of 25%, while those who spun a 65 provided an average guess of 45%. They rationalized these results by explaining that people make estimates by starting from an initial value, and their adjustments from this quantity are typically insufficient.

While initial experiments investigating the anchoring effect recruited amateur participants,[19] researchers also observed similar anchoring effects among experts. In their seminal study from 1987, Gregory Northcraft and Margaret Neale recruited real estate agents to visit a home, review a detailed booklet containing information about the property, and then assess the value of the house.[16] The researchers listed a low asking price in the booklet for one group (low anchor) and a high asking price for another group (high anchor). The agents who viewed the high asking price provided valuations 41% greater than those who viewed the lower price, and the anchoring index of the listing price was likewise 41%. Northcraft and Neale conducted an identical experiment among business school students with no real estate experience and observed similar results: the students in the high anchor treatment answered with valuations that exceeded those in the low anchor treatment by 48%, and the anchoring index of the listing price was also 48%. Their findings, therefore, suggested that anchors such as listing prices bias the decisions of trained professionals and inexperienced individuals similarly.

Even if algorithms do not officially make decisions, they anchor human decisions in serious ways.

More recent research finds evidence of the anchoring effect in the criminal justice system. In 2006, Birte Englich, Thomas Mussweiler, and Fritz Strack conducted a study in which judges threw a pair of dice and then provided a prison sentence for an individual convicted of shoplifting.[7] The researchers rigged the dice so they would land on a low number (low anchor) for half of the participants and a high number (high anchor) for the other half. The judges who rolled a low number provided an average sentence of five months, whereas the judges who rolled a high number provided an average sentence of eight months. The difference in responses was statistically significant, and the anchoring index of the dice roll was 67%. In fact, similar studies have shown that sentencing demands,[7] motions to dismiss,[13] and damages caps[15] also act as anchors that bias judges' decision-making.

Methods. This second experiment thus sought to investigate if algorithmic risk scores influence human decisions by serving as anchors. The experiment entailed a 1 x 2 between-subjects design where the two treatments were as follows: low score, in which participants viewed the defendant profile accompanied by a low-risk score; and high-score, in which participants viewed the defendant profile accompanied by a high-risk score.

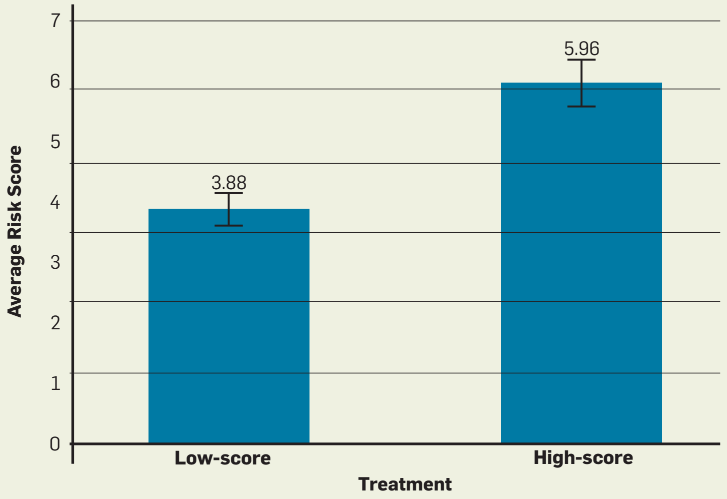
The low-score and high-score treatments assigned risk scores based on the original COMPAS score according to the following formulas:

Low-score = max(0,COMPAS − 3)
High-score = min(10,COMPAS + 3)

This new experiment mirrored the previous one: Participants evaluated the same 40 defendants, met the same requirements, and received the same payment. The study also employed the format on the Qualtrics platform.

Results. Figure 4 shows the average scores of participants assigned to defendants versus those provided in the defendant profiles in the low-score and high-score treatments. Error bars represent the 95% confidence intervals. The scores that participants assigned defendants highly correlate with those that they viewed in the defendants' profile descriptions. As such, participants in the low-score treatment provided risk scores that were, on average, 42.3% lower than participants in the high-score treatment when assessing the same set of defendants. The average risk score from respondents in the low-score treatment was 3.88 (95% CI 3.39–4.36), while the average risk score from respondents in the high-score treatment was 5.96 (95% CI 5.36–6.56). A two-sided *t*-test revealed that this difference was statistically significant ($p < 0.0001$).

Figure 4. Average risk score by treatment.



At the end of the survey, when participants reflected on the role of the COMPAS algorithm in their decision-making, they indicated common themes, such as using the algorithm's score as a starting point and as a verification of their own decisions. The table in Figure 5 summarizes these participant comments by their treatment group and role of the algorithm in their decision-making.

Discussion. The results from this study indicate that algorithmic risk predictions serve as anchors that bias human decision-making. Participants in the low-score treatment provided an average risk score of 3.88, while participants in the high-score treatment assigned an average risk score of 5.96. The average anchoring index across all 40 defendants was 56.71%. This anchor measure mirrored that found in prior psychology literature.[8,14,16] For example, one study investigated the anchoring bias in estimations by asking participants to guess the height of the tallest redwood tree.[14] The researchers provided one group with a low anchor of 180 feet and another group with a high anchor of 1,200 feet, and they observed an anchoring index of 55%. Scholars have observed similar values of the anchoring index in contexts such as probability estimates,[19] purchasing decisions,[20] and sales forecasting.[5]

Even though this type of cognitive bias occurs among participants with little training in the criminal justice system, prior work suggests the anchoring effect varies little between non-experts and experts in a given field. Northcraft and Neale found that asking prices for homes similarly influenced real estate agents and people with no real estate experience.[16] This study thus suggested that the anchoring effect of algorithmic risk assessments among judges, bail, and parole officers would mirror that of the participants in this experiment. Numerous prior studies demonstrate that these officials are, in fact,

Figure 5. Responses by treatment group and algorithm role.

Algorithm Role	Low Treatment	High Treatment
Factor	"I took them into consideration but still made my own decisions."	"I took it into account, but did not count on it 100 percent." "I used it as *one* factor."
Tipping Point	"I used it if I was wavering on a score." "I would look at it if I was close on which way to go."	"For those cases where I felt a 50/50 chance, I sided with the COMPAS score."
Validation	"Only considered it when it seemed to coincide with my own judgment."	"I looked to see if it was similar to what I thought."
Guideline	"I used it to target my general range of scores, unless I had reason to strongly disagree." "I used the scores to base the start value of my score, read the description of their crime and modified the score."	"I kind of started with the COMPAS risk score, and then raised or lowered the score based on previous criminal history (or lack of one)." "I used it as a guideline to structure my decisions on."
Deference	"I always considered it as near perfect"	NA
Ignored	"I usually ignored it because it didn't seem like it made much sense to me, but who knows." "It didn't seem very consistent or accurate so I didn't factor it in much, if at all."	"I *hate* rubrics. You are looking at people, dynamic people, and a computerized rubric or other type of system designed to assess risks is completely ignoring so many other *very important* circumstances that may affect these odds."

susceptible to forms of cognitive bias such as anchoring.[7,15]

These findings also, importantly, highlight problems with existing frameworks to address machine bias. For example, many researchers advocate for putting a "human in the loop" to act in a supervisory capacity, and they claim this measure will improve accuracy and, in the context of risk assessments, "ensure a sentence is just and reasonable."[12] Even when humans make the final decisions, however, the machine-learning models exert influence by anchoring these decisions. An algorithm's output still shapes the ultimate treatment for defendants.

The subtle influence of algorithms via this type of cognitive bias may extend to other domains such as finance, hiring, and medicine. Future work should, no doubt, focus on the collaborative potential of humans and machines, as well as steps to promote algorithmic fairness. But this work must consider the susceptibility of humans when developing measures to address the shortcomings of machine learning models.

Conclusion

The COMPAS algorithm was used here as a case study to investigate the role of algorithmic risk assessments in human decision-making. Prior work on the COMPAS algorithm and similar risk-assessment instruments focused on the technical aspects of the tools by presenting methods to improve their accuracy and theorizing frameworks to evaluate the fairness of their predictions. The research has not considered the practical function of the algorithm as a decision-making aid rather than as a decision maker.

Based on the theoretical findings from the existing literature, some policymakers and software engineers contend that algorithmic risk assessments such as the COMPAS software can alleviate the incarceration epidemic and the occurrence of violent crimes by informing and improving decisions about policing, treatment, and sentencing.

The first experiment described here thus explored how the COMPAS algorithm affects accuracy in a controlled environment with human subjects. When predicting the risk that a defendant will recidivate, the COMPAS algorithm achieved a significantly higher accuracy rate than the participants who assessed defendant profiles (65.0% vs. 54.2%). Yet when participants incorporated the algorithm's risk assessments into their decisions, their accuracy did not improve. The experiment also evaluated the effect of presenting an advisement designed to warn of the potential for disparate impact on minorities. The findings suggest, however, that the advisement did not significantly impact the accuracy of recidivism predictions.

Moreover, researchers have increasingly devoted attention to the fairness of risk-assessment software. While many people acknowledge the potential for algorithmic bias in these tools, they contend that leaving a human in the loop can ensure fair treatment for defendants. The results from the second experiment, however, indicate that the algorithmic risk scores acted as anchors that induced a cognitive bias: Participants assimilated their predictions to the algorithm's score. Participants who viewed the set of low-risk scores provided risk scores, on average, 42.3% lower than participants who viewed the high-risk scores when assessing the same set of defendants. Given this human susceptibility, an inaccurate algorithm may still result in erroneous decisions.

Considered in tandem, these findings indicate that collaboration between humans and machines does not necessarily lead to better outcomes, and human supervision does not sufficiently address problems when algorithms err or demonstrate concerning biases. If machines are to improve outcomes in the criminal justice system and beyond, future research must further investigate their practical role: an input to human decision makers.

Related articles on queue.acm.org

The Mythos of Model Interpretability
Zachary C. Lipton
https://queue.acm.org/detail.cfm?id=3241340

The API Performance Contract
Robert F. Sproull and Jim Waldo
https://queue.acm.org/detail.cfm?id=2576968

Accountability in Algorithmic Decision-Making
Nicholas Diakopoulos
https://queue.acm.org/detail.cfm?id=2886105

References

1. Angwin, J., Larson, J. Machine bias. *ProPublica* (May 23, 2016).
2. Case, N. How to become a centaur. *J. Design and Science* (Jan. 2018).
3. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data 5*, 2 (2017), 153–163.
4. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM Press, 2017, 797–806.
5. Critcher, C.R. and Gilovich, T. Incidental environmental anchors. *J. Behavioral Decision Making 21*, 3 (2008), 241–251.
6. Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*, 1 (2018), eaao5580.
7. Englich, B., Mussweiler, T. and Strack, F. Playing dice with criminal sentences: the influence of irrelevant anchors on experts' judicial decision making. *Personality and Social Psychology Bulletin 32*, 2 (2006), 188–200.
8. Furnham, A. and Boo, H.C. A literature review of the anchoring effect. *The J. Socio-Economics 40*, 1 (2011), 35–42.
9. Goldstein, I.M., Lawrence, J. and Miner, A.S. Human-machine collaboration in cancer and beyond: The Centaur Care Model. *JAMA Oncology 3*, 10 (2017), 1303.
10. Green, K.C. and Armstrong, J.S. Evidence on the effects of mandatory disclaimers in advertising. *J. Public Policy & Marketing 31*, 2 (2012), 293–304.
11. Horvitz, E. and Paek, T. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction 17*, 1-2 (2007), 159–182.
12. Johnson, R.C. Overcoming AI bias with AI fairness. *Commun. ACM* (Dec. 6, 2018).
13. Jukier, R. Inside the judicial mind: exploring judicial methodology in the mixed legal system of Quebec. *European J. Comparative Law and Governance* (Feb. 2014).
14. Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.
15. Mussweiler, T. and Strack, F. Numeric judgments under uncertainty: the role of knowledge in anchoring. *J. Experimental Social Psychology 36*, 5 (2000), 495–518.
16. Northcraft, G.B. and Neale, M.A. Experts, amateurs, and real estate: an anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes 39*, 1 (1987), 84–97.
17. Shaw, A.D., Horton, J.J. and Chen, D.L. Designing incentives for inexpert human raters. In *Proceedings of the ACM Conf. Computer-supported Cooperative Work*. ACM Press, 2011, 275–284.
18. *State v Loomis*, 2016.
19. Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science 185*, 4157 (1974), 1124–1131.
20. Wansink, B., Kent, R.J. and Hoch, S.J. An anchoring and adjustment model of purchase quantity decisions. *J. Marketing Research 35*, 1 (1998), 71.

Michelle Vaccaro received a bachelor's degree in computer science in 2019 from Harvard College, Cambridge, MA, USA.

Jim Waldo is a Gordon McKay Professor of the practice of computer science at Harvard University, Cambridge, MA, USA, where he is also a professor of technology policy at the Harvard Kennedy School. Prior to joining Harvard, he spent more than 30 years in the industry, much of that at Sun Microsystems.

DOI:10.1145/3359334

Article development led by acmqueue
queue.acm.org

The trade-offs between write and read.

BY PAT HELLAND

Write Amplification vs. Read Perspiration

INCREASINGLY IN COMPUTING systems, when you write something into durable storage it is in need of reorganization later. Personally, I'm pretty darned disorganized and I lose stuff a lot. This causes extensive searching, sometimes to no avail. It is, however, easier to "store" stuff by setting it down wherever I feel like it.

In computing, there is an interesting trend where writing creates a need to do more work. You need to reorganize, merge, reindex, and more to make the stuff you wrote more useful. If you don't, you must search or do other work to support future reads.

Indexing within a database. My first programming job was to implement a database system. In 1978, my colleague and I didn't even know what that was! We voraciously read every paper from ACM's Special Interest Group on Management of Data and *ACM Transactions on Database Systems* we could lay our hands on. We learned about this interesting and confusing concept of a relational database and how indexing can optimize access while being transparent to the application. Of course, updating an index meant another two-disk access since the indices of a B+ tree didn't fit in memory. We understood the additional work to make database changes was worth it if you were ever going to read it later.

The next perplexing question was: How much should be indexed? Should we index every column? When should a pair of columns be indexed together? The more indexing we did, the faster the read queries would become. The more indexing we did, the more our ability to update became slower than molasses.

I learned this is a common trade-off. Reading fast frequently means writing slow.

Row-store vs. column-store. I have focused most of my misspent career on distributed systems and online transaction processing (OLTP)-style databases. It's natural for me to associate high-performance updates with what today is called a *row-store*.

Another approach is to organize data by columns: Take a bunch of rows and organize the data by its column values. Every row containing the state of California, for example, keeps just the single column's data together. Columnar databases are super fast for doing queries because many logical rows with the same value are physically close to each other.

However, updating a *column-store* is not as easy. Typically, updates are kept separately in an integrated row-store. Queries check the small row-store in a fashion that's relatively fast because it's small. These queries are combined with the results of the faster column-store to give a unified accurate answer. Periodically, the new row-store updates are merged with the column-store to make a new column-store. This may be done in a cascading fashion somewhat like the merges in an log-structured merge (LSM) tree, described in the next section.

When inserting into a column-store (or really its attached row-store), you are incurring a debt to be paid later. This debt to rewrite and integrate the new data is a form of *write amplification* where a single write turns into more writes later.

LSM trees were first proposed in 1996.[6] The idea is to track changes to a key-value store as transactions, with new values kept in memory. As transactions commit, the sorted collection of recent key-value pairs can be written to disk in a uniquely named file. This file contains the sorted key-value pairs along with an index into the keys in the file. Once written to disk, the newly committed changes do not need to be kept in memory.

Now, if you keep doing this, looking up values by key starts looking like what happens to me when I try to find something I set down in some random place. Linear searches for your wallet might be tractable in a small apartment but not so much when the search space gets bigger in a larger home in the suburbs. To reduce the *read perspiration*, LSM trees invest energy to organize the data by rewriting it as you go.

Search makes reading the documents a lot easier. It dramatically lowers the read perspiration.

When a new file is freshly written from the storage engine, it has a bunch of key-value pairs. To make it easy to find keys, these are merged with files that were written earlier. Each LSM tree has some form of fan-out where lower levels of the tree (with keys written earlier) are kept across more files. For example, you may have 10 times as many files at level 1 as at the brand-new level 0. Each file at level 1 has approximately one-tenth as large a key range represented but approximately 10 times the amount of update time represented. Similarly, moving down to level 2 results in 100 files, each with a narrower key range and longer time range.

The depth of an LSM tree depends on the fan-out, the size of each file, and the number of key-value pairs in the tree. In general, most of the storage is in the lowest level of the tree.

So, within this basic LSM structure that is gaining so much popularity, there are varieties of implementation choices. Consider:

- *Leveling merges.* When a new file is added to a level, pick the next file in the round-robin traversal and merge it with the files in the next level below. Suppose you pick a fan-out of 10; you will find the key range in the file dropping down typically covers the key range in about 10 files in the level below. You merge 11 files together as one drops down onto 10 and you get 11 files out. Now, the next level has gotten fatter by one file, so you repeat and merge down again.
- *Tiering merges.* In this different but related approach, you let a bunch of files stack up on each level before doing the merge. Say you stack up 10 files before you merge down at each level. That dramatically reduces the amount of merging required.

Leveling merges have a large write amplification. Each write of a new key-value pair to level 0 will be rewritten 10 or 11 times at each level it moves through. On the other hand, they have a small read perspiration, as a reader typically checks only one place per level.

Tiering merges have a much lower write amplification but a larger read perspiration. Because new files stack up at each level before merging, there is less merging and hence less writing. On the other hand, reads must check a lot more places, leading to the larger read perspiration.

There's a bunch of fun work lately on the trade-offs of these schemes.[2,5]

Indexing and searching. Search is in many ways a variation of database indexing. In database indices, the notion of identity exists hidden within the database as a row-id or a primary key. Within a relational system, updates to indices are transactionally integrated, and the user sees only a performance difference.

Search systems are a bit different in that they deal with documents. Most search systems asynchronously update the search index *after* the change to the document occurs. This is knit together with some form of document identity.[3]

Search makes reading the documents a lot easier. It dramatically lowers the read perspiration. Updates to the documents asynchronously impose a debt onto the system to get them indexed. Creating and merging search indices is a complex job that I think of as a form of write amplification.

To index, you must scour the corpus to find recently written or updated documents. Each of these needs to have an identifier and then must be processed to locate the search terms (sometimes called n-grams; https://en.wikipedia.org/wiki/n-gram). Each of these many n-grams found in a typical document then needs to be sent to an indexer that covers one of many shards. So, the document identifier now is associated with each term (or n-gram) located in the searchable document—all of this because the user did a write or created a document!

I worked for a few years on an Internet-scale search engine and know how they work. I'm still in awe that all this machinery can keep up with the work involved in all that write amplification. It's a lot of work for each document written—and there are lots and lots of documents.

Internet-scale search systems clearly offer excellent and low read perspiration.

Large-scale caches. Lots of big Internet systems have ginormous caches. Consider a product catalog at a big ecommerce retailer. Whenever anything changes, lots of servers are updated with the new product description. This makes for a very easy and fast read in exchange for a lot of writes.

Normalization and denormalization. Growing up in the relational database world, I was imbued with the determination to have normalized data contained in the database. Working to avoid update anomalies was deemed to be extremely important. Performing a large number of joins to get an answer was a small penalty to pay to ensure the database wasn't damaged by an errant update.

Increasingly, I view this as the equivalent of throwing salt over your shoulder if you spill some. Yeah… I've seen others do it, but I'm not sure I should.

Most systems are getting more distributed. Most of these have key-value pairs containing their data, which is sharded for scale. By grouping related data into the value of a pair—typically in a JSON (JavaScript Object Notation) representation or something similar—it's easy to grab the value, perhaps as a string, and squirt it over to the distant system issuing the request.

If you *were* to normalize the data in this big and sharded system, the normalized values would not be on the same shard together. Doing a distributed join is more annoying than doing a centralized join.

To cope with this, people superimpose versioning on their data. It's not perfect but it's less challenging than distributed joins or trying to do massive updates across the denormalized data. The classic example for the value of normalization in databases is a denormalized table with employees, their manager, and their manager's phone number.[4] Because the manager's phone number is copied in many tables for many employees, it's hard to change it. Increasingly, I see systems store "as-of" data in their denormalized structures—for example, the manager's phone is captured "as-of" June 1.

Large-scale distributed systems put a lot of pressure on the semantics of a consistent read. This, in turn, can be seen as a tension between write amplification and read perspiration.

Conclusion

I have looked at just a few of the examples where there are trade-offs in our systems between write and read.[1] It is endemic in so many environments. We see emerging systems that adapt and optimize for these trade-offs as they watch their usage patterns. Fun stuff! C

Related articles on queue.acm.org

Immutability Changes Everything
Pat Helland
https://queue.acm.org/detail.cfm?id=2884038

Disambiguating Databases
Rick Richardson
https://queue.acm.org/detail.cfm?id=2696453

The Pathologies of Big Data
Adam Jacobs
https://queue.acm.org/detail.cfm?id=1563874

References

1. Athanassoulis, M., Kester, M.S., Maas, L. M., Stoica, R., Idreos, S., Ailamaki, A. and Callaghan, M. Designing access methods: The RUM conjecture. In *Proceedings of the 19th International Conference on Extending Database Technology* (2016).
2. Dayan, N. and Idreos, S. Dostoevsky: better space-time tradeoffs for LSM-tree-based key-value stores via adaptive removal of superfluous merging. In *Proceedings of the Intern. Conf. Management of Data* (2018), 505–520.
3. Helland, P. Identity by any other name. *Commun. ACM 62*, 4 (Apr. 2019), 80.
4. Helland, P. Normalization is for sissies (July 23, 2007); http://bit.ly/30iL7g3
5. Luo, C., and Carey, M.J. Forthcoming. LSM-based storage techniques. *Computing Surveys*; arXiv:1812.07527.
6. O'Neil, P., Cheng, E., Gawlick, D. and O'Neil, E. The log-structured merge-tree (LSM-tree). *Acta Informatica 33*, 4 (1996).

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.

DOI:10.1145/3318163

Tracing the evolution of the five-minute rule to help identify imminent changes in the design of data management engines.

BY RAJA APPUSWAMY, GOETZ GRAEFE, RENATA BOROVICA-GAJIC, AND ANASTASIA AILAMAKI

The Five-Minute Rule 30 Years Later and Its Impact on the Storage Hierarchy

THE DESIGN OF data management systems has always been influenced by the storage hardware landscape. In the 1980s, database engines used a two-tier storage hierarchy consisting of dynamic random access memory (DRAM) and hard disk drives (HDD). Given the disparity in cost between HDD and DRAM, it was important to determine when it made economic sense to cache data in DRAM as opposed to leaving it on the HDD.

In 1987, Jim Gray and Gianfranco Putzolu established the five-minute rule that gave a precise answer to this question: "1KB records referenced every five minutes should be memory resident."[9] They arrived at this value by using the then-current price-performance characteristics of DRAM and HDD shown in Table 1 for computing the break-even interval at which the cost of holding 1KB of data in DRAM matches the cost of I/O to fetch it from HDD.

Today, enterprise database engines use a three-tier storage hierarchy as depicted in Figure 1. DRAM or NAND flash solid state device (SSD)-based performance tier is used for hosting data accessed by latency-critical transaction processing and real-time analytics applications. The HDD-based capacity tier hosts data accessed by latency-insensitive batch analytics applications. The archival tier is not used for online query processing, but for storing data that is only accessed rarely during regulatory compliance audits or disaster recovery. This tier is primarily based on tape and is extremely crucial as a long-term data repository for several application domains like physics, banking, security, and law enforcement.

In this article, we revisit the five-minute rule three decades after its inception. We recomputed break-even intervals for each tier of the modern, multi-tiered storage hierarchy and use guidelines provided by the five-minute rule to identify impending changes in the design of data management engines for emerging storage hardware. We summarize our findings here:

- **HDD is tape.** The gap between DRAM and HDD is increasing as the five-minute rule valid for the DRAM–HDD case in 1987 is now a four-hour rule. This implies the HDD-based capacity tier is losing relevance for not just performance sensitive applications, but for all applications with a non-sequential data access pattern.
- **Non-volatile memory is DRAM.** The gap between DRAM and SSD is shrinking. The original five-minute rule is now valid for the DRAM–SSD case, and the break-even interval is less than a minute for newer non-volatile memory (NVM) devices like 3D-

XPoint.[23] This suggests an impending shift from DRAM- based database engines to flash or NVM-based persistent memory engines.

▸ **Cold storage is hot.** The gap between HDD and tape is also rapidly shrinking for sequential workloads. New cold storage devices that are touted to offer second-long access latency with cost comparable to tape reduce this gap further. This suggests the HDD-based capacity tier will soon lose relevance even for non-performance-critical batch analytics applications that can be scheduled to run directly over newer cold storage devices.

Revisiting the Five-Minute Rule

The five-minute rule explores the trade-off between the cost of DRAM and the cost of disk I/O by providing a formula to predict the break-even interval—the time window within which data must be reaccessed in order for it to be economically beneficial to be cached in DRAM. The interval is computed as:

$$\frac{\text{PagesPerMBofDRAM}}{(\text{AccessesPerSecondPerDisk})} \times \frac{\text{PricePerDiskDrive}}{\text{PricePerMBofDRAM}} \quad (1)$$

The first ratio in the equation was referred to as the technology ratio, as random I/O access capability of the secondary storage device, and the page size used by the database engine for performing I/O, both directly depend on the hardware technology used for secondary storage. The second ratio, in contrast, is referred to as the economic ratio as pricing is determined by factors other than just hardware technology. Rearranging the formulation by swapping the denominators provides the intuition behind the five-minute rule, as it reduces the equation to price-per-disk-access-per-second normalized by the price-per-page of DRAM. This term directly compares the cost of performing I/O to fetch a page from disk versus the cost of caching it in DRAM.

Table 1 shows the price, capacity, and performance of DRAM, HDD, and NAND flash-based SSDs across four decades. The values shown for 1987, 1997, and 2007 are those reported by previous revisions of the five-minute rule.[6,8,9] The values listed for 2018 are performance metrics listed in vendor specifications, and unit price quoted by www.newegg.com as of Mar. 1, 2018, for DRAM, SSD, and HDD components specified in a recent TPC-C report.[24]

DRAM–HDD. Table 2 presents both the break-even interval for 4KB pages and the page sizes for which the five-minute rule is applicable across four decades. In 1987, the break-even inter-

IMAGE BY DEVENORR

Table 1. The evolution of DRAM, HDD, and Flash SSD properties.

Metric	DRAM				HDD				SATAFlash SSD	
	1987	1997	2007	2018	1987	1997	2007	2018	2007	2018
Unit price($)	5k	15k	48	80	30k	2k	80	49	1k	415
Unit capacity	1MB	1GB	1GB	16GB	180MB	9GB	250GB	2TB	32GB	800GB
$/MB	5k	14.6	0.05	0.005	83.33	0.22	0.0003	0.00002	0.03	0.0005
Random IOPS	–	–	–	–	5	64	83	200	6.2k	67k (r)/20k (w)
Sequential b/w (MB/s)	–	–	–	–	1	10	300	200	66	500 (r)/460 (w)

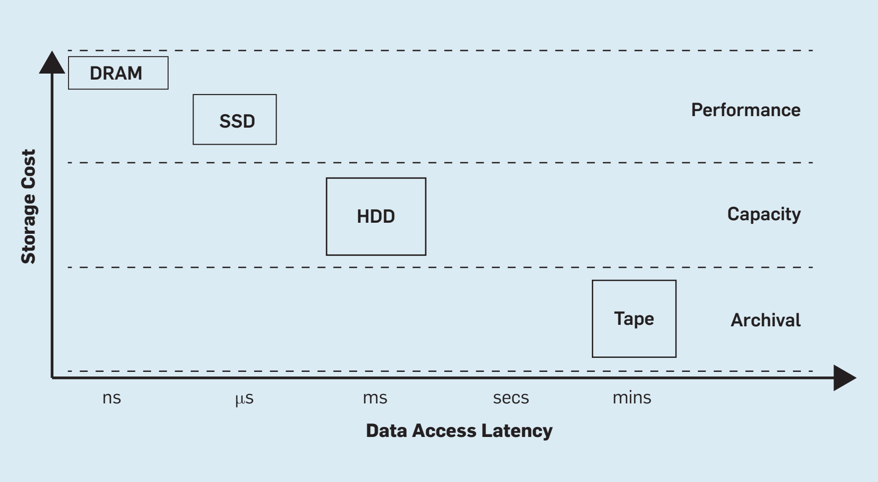
val was 400 seconds for 1KB pages. This was rounded down to five minutes, thus, lending the name for the rule. For 4KB pages, the break-even interval was 100 seconds. When the study was repeated in 1997, the break-even interval had increased to nine minutes for 4KB pages, and the five-minute rule was determined to hold only for 8KB pages. Between 1997 and 2007, DRAM and HDD prices dropped further resulting in the economic ratio increasing from 133 ($2k/$15) to 1700 ($80/$0.05). However, the technology ratio did not drop proportionately due to a lack of improvement in HDD random access latency. As a result, the break-even interval for 4KB pages increased 10×, from nine minutes to 1.5 hours. The five-minute rule was applicable only for 64KB pages in 2007.

Continuing this trend, the break-even interval for DRAM–HDD case today is four hours for 4KB pages. The five-minute rule is valid today for 512KB pages. The break-even interval trend indicates it is more economical to store most data in DRAM instead of the HDD, and the page size trend indicates that even rare accesses to HDD should be performed in large granularities.

DRAM–SSD. SSDs are being increasingly used as the storage medium of choice in the latency-critical performance tier due to their superior random access capability compared to HDDs. Thus, the five-minute rule can be used to compute a break-even interval for the case where DRAM is used to cache data stored in SSDs. Table 3 shows the interval in 2007, when SSDs were in the initial stages of adoption, and today, based on metrics listed in Table 1.

We see the interval has dropped from 15 minutes to five minutes for 4KB pages. Thus, the five-minute rule is valid for SSDs today. This is in stark contrast with the DRAM–HDD case, where the interval increased 2.7× from 1.5 hours to four hours. In both DRAM–HDD and DRAM–SSD cases, the drop in DRAM cost/MB dominated the economic ratio. However, unlike the 2.5× improvement in random I/Os-per-second (IOPS) with HDDs, SSDs have managed to achieve an impressive 11× improvement (67k/6.2k). Thus, the increase in economic ratio was overshadowed by the decrease in technology ratio with SSDs, resulting in the interval shrinking.

Figure 1. Storage tiering for enterprise databases.



SSD–HDD. As SSDs can also be used as a cache for HDD, the same formula can also be used to estimate the break-even interval for the SSD–HDD case. From Table 3, we see the break-even interval for this case has increased by a factor of 10× from 2.25 hours in 2007 to 1.5 days in 2018. The SSD–HDD interval is nine times longer than the DRAM–HDD interval of four hours.

Implications. There are two important consequences of these results. First, in 2007, the turnover time in the DRAM–HDD case was six times higher than the DRAM–SSD case (1.5h/15m). In 2018, it is nearly 50× higher (4h/5m). Thus, in systems tuned using economic considerations, one should replace HDD with SSD, as it would not only improve performance, but also reduce the amount of DRAM required for caching data. Second, given the four-hour DRAM–HDD and one day SSD–HDD intervals, it is important to keep all active data in the DRAM or SSD-based performance tier and relegate the HDD-based capacity tier to storing only infrequently accessed data. The growing gap between performance and capacity tiers also implies that SSD vendors should optimize for $/IOPS, and HDD vendors, in contrast, should optimize for $/GB. Next, we highlight recent changes in performance and capacity tiers that indicate such targeted optimizations are already underway.

The Performance Tier

NAND flash. NAND flash-based solid-state storage has been steadily inching its way closer to the CPU over the past two decades. When NAND flash was introduced in the early 2000s, solid-state storage was dominated by

DRAM-based SSD products. By the mid 2000s, improvements in performance and reliability of NAND flash resulted in flash-based serial AT attachment (SATA) SSDs gaining popularity in niche application domains. The late 2000s witnessed the emergence of a new breed of peripheral component interconnect express (PCIe) flash SSDs that could deliver two orders of magnitude higher throughput than their SATA counterparts. Since then, a rapid increase in capacity, drop in pricing, and new low-overhead interfaces like non-volatile memory express (NVMe), have all resulted in PCIe flash SSDs displacing their SATA counterparts as server accelerators of choice.

Table 4 (first row) shows the price/performance characteristics of a representative, state-of-the-art PCIe SSD. In comparison to Table 1, we find the PCIe SSD offers five times higher read IOPS and sequential access bandwidth than its SATA counterpart.

NVDIMM. As SSD vendors continue to improve throughput and capacity, the bottleneck in the storage subsystem has shifted from the device itself to the PCIe bus that is used to interface with the SSD. Thus, over the past few years, NAND flash has started transitioning once again from storage devices that are interfaced via the high-latency, bandwidth-limited PCIe bus into non-volatile memory (NVM) devices that are interfaced via the low-latency, high-bandwidth memory bus. These devices, also referred to as non-volatile DIMMs (NVDIMM), use a combination of DRAM and flash storage media packaged together as a dual in-line memory module (DIMM).

NVM. Today, NVDIMMs are niche accelerators compared to PCIe SSDs due to a high cost/GB. Unlike these NVDIMM technologies that rely on NAND flash, new NVM technologies that are touted to have better endurance, higher throughput, and lower latency than NAND flash are being actively developed.

Table 4 (second row) shows the characteristics of Intel Optane DC P4800X—a PCIe SSD based on 3D XPoint, a new phase-changed-media-based NVM technology. The cost/GB of 3D XPoint is higher than NAND flash today as the technology is yet to mature. However, preliminary studies have found that 3D XPoint provides predictable access latencies that are much lower than several state-of-the-art NAND flash devices even under severe load.[23]

Break-even interval and implications. When we apply the five-minute rule formula using metrics given in Table 4, we get a break-even interval of one minute for 4KB pages in both the DRAM–NAND Flash PCIe SSD and DRAM–3D XPoint cases. Comparing these results with Table 2, we see that the breakeven interval is 10× shorter when PCIe SSDs or new PM technologies are used as the second tier instead of SATA SSDs. This can be attributed to the drop in technology ratio caused by the improvement in random IOPS.

Implications. Today, in the era of in-memory data management, several database engines are designed based on the assumption that all data is resident in DRAM. However, the dramatic drop in breakeven interval computed by the five-minute rule challenges this trend of DRAM-based in-memory data management due to three reasons. First, recent projections indicate that flash density is expected to increase 40% annually over the next five years.[5] DRAM, in contrast, is doubling in capacity every three years.[17] As a result, the cost of NAND flash is likely to drop faster than DRAM. This, in turn, will result in the economic ratio dropping further leading to a reduction in the break-even interval.

Second, modern PCIe SSD is a highly parallel device that can provide very high random I/O throughput by servicing multiple outstanding I/Os concurrently. New non-volatile memory technologies like 3D XPoint promise further improvements in both throughput and access latencies over NAND flash. With interfaces like NVMe, the end-to-end latency of accessing data from PCIe 3D XPoint SSDs is just tens of µs. Thus, further improvements in non-volatile solid-state storage media will result in a drop in technology ratio, thereby reducing the break-even interval further.

Third, SSDs consume substantially lower power than DRAM. The Intel 750 SSD consumes 4W of power when idle and 22W when active. In contrast, 1TB of DRAM in a server would consume 50W when idle and 100W when active.[1] It is also well known that DRAM power consumption increases non-linearly with capacity, as high-density DRAM consumes substantially more power than their low-density counterparts. A recent study that focuses on power

Table 2. The evolution of the page size for which the five-minute rule holds across four decades based on appropriate price, performance, and page size values.

	1987	1997	2007	2018
Break-even (4KB page)	100s	9m	1.5h	4h
Page size (5-minute interval)	1KB	8KB	64KB	512KB

Table 3. The evolution of the break-even interval across four decades based on appropriate price, performance, and page size values.

Tier	1987	1997	2007	2018
DRAM–SSD	—	—	15m	5m
SSD–HDD	—	—	2.25h	1.5d

Table 4. Price/performance metrics for the NAND-based Intel 750 PCIe SSD and 3D-XPoint-based Intel Optane P4800X PCIe SSD.

Device	Capacity	Price($)	IOPS(k)	B/w(GB/s)
Intel 750	800GB	589	460	2.5
Intel P4800X	480GB	617	550	2.5

consumption in main memory databases showed that in a server equipped with 6TB of memory, the idle power of DRAM would match that of four active CPUs.[1] Such a difference in power consumption between SSD and DRAM directly translates into higher Operational Expenses (OPEX), and hence, higher Total Cost of Ownership (TCO), for DRAM-based database engines.

Given these three factors, the break-even interval from the five-minute rule seems to suggest an inevitable shift from DRAM-based data management engines to NVM-based persistent-memory engines. In fact, this change is already well under way, as state-of- the-art database engines are being updated to fully exploit the performance benefits of PCIe NVMe SSDs.[26] Researchers have recently highlighted the fact that data caching systems that trade-off performance for price by reducing the amount of DRAM are gaining market share over in-memory database engines.[18]

The Capacity Tier

HDD. Traditionally, HDDs have been the primary storage media used for provisioning the capacity tier. For several years, areal density improvements enabled HDDs to increase capacity at Kryder's rate (40% per year), outstripping Moore's Law. However, over the past few years, HDD vendors have hit walls in scaling areal density with conventional Perpendicular Magnetic Recording (PMR) techniques resulting in annual areal density improvement of only around 16% instead of 40%.[19]

HDDs also present another problem when used as the storage medium of choice for building a capacity tier, namely, high idle power consumption. Although enterprises gather vast amounts of data, as one might expect, not all data is accessed frequently. Recent studies estimate that as much as 80% of enterprise data is "cold," meaning infrequently accessed, and that cold data is the largest growing segment with a 60% Cumulative Annual Growth Rate (CAGR).[10–12] Unlike tape, which consumes no power once unmounted, HDDs consume a substantial amount of power even while idle. Such power consumption translates to a proportional increase in TCO.

Tape. The areal density of tape has been increasing steadily at a rate of 33% per year and roadmaps from the Linear Tape Open consortium (LTO)[25] and the Information Storage Industry Consortium (INSIC)[13] project a continued increase in density for the foreseeable future.

Table 5. Price/performance characteristics of tape.

	1997	2018
Tape library cost ($)	10,000	11,000
Number of drives	1	4
Number of slots	14	10
Max capacity per tape	35GB	15TB
Transfer rate per drive (MB/s)	5	750
Access latency	30s	65s

Table 6. Price/performance metrics of DRAM, HDD, and tape.

Metric	DRAM	HDD	Tape
Unit capacity	16GB	2TB	10 × 15TB
Unit cost ($)	80	50	11,000
Latency	100ns	5ms	65s
Bandwidth	100 GB/s	200 MB/s	4 × 750MB/s
Kaps	9,000,000	200	0.02
Maps	10,000	100	0.02
Scan time	0.16s	3hours	14hours
$/Kaps	9e-14	5e-09	8e-03
$/Maps	9e-12	8e-09	8e-03
$/Tbscan	8e-06	0.003	0.03
$/TBscan (97)	0.32	4.23	296

Table 5 shows the price/performance metrics of tape storage both in 1997 and today. The 1997 values are based on the corresponding five-minute rule paper.[8] The 2018 values are based on a SpectraLogic T50e tape library[22] using LTO-7 tape cartridges.

With individual tape capacity increasing 200× since 1997, the total capacity stored in tape libraries has expanded from hundreds of gigabytes to hundreds of petabytes today. Further, a single LTO-7 cartridge is capable of matching, or even outperforming a HDD, with respect to sequential data access bandwidth as shown in Table 6. As modern tape libraries use multiple drives, the cumulative bandwidth achievable using even low-end tape libraries is 1–2GB/s. High-end libraries can deliver well over 40GB/s. These benefits have made tape the preferable media of choice in the archival tier both on-premise and in the cloud, for several applications ranging from natural sciences, like particle physics and astronomy, to movies archives in the entertainment industry.[15,20] However, random access latency of tape is still 1000× higher than HDD (minutes vs. ms) due to the fact that tape libraries need to mechanically load and wind tape cartridges before data can be accessed.

Break-even interval and implications. Using metrics from Tables 1, 5 to compute the break-even interval for the DRAM–tape case results in an interval of over 300 years for a page size of 4KB! Jim Gray referred to tape drives as the "data motel" where data checks in and never checks out,[7] and this is certainly true today. Figure 2 shows the variation in break-even interval for both HDD and tape for various page sizes. We see that the interval asymptotically approaches one minute in the DRAM–HDD case and 10 minutes in the DRAM–tape case. The HDD asymptote is reached at a page size of 100MB and the tape asymptote is reached at a size of 100GB. This clearly shows that randomly accessing data on these devices is extremely expensive, and data transfer sizes with these devices should be large to amortize the cost of random accesses.

However, the primary use of the capacity tier today is not sup-porting applications that require high-performance random accesses. Rather, it is to reduce the cost/GB of storing data over which latency-insensitive batch analytics can be performed. Indeed, Gray and Graefe noted that metrics like KB-accesses-per-second (Kaps) are less relevant for HDD and tape as they grow into infinite-capacity resources.[8] Instead, MB-accesses-per-second (Maps) and time to scan the whole devices are more pertinent to these high-density storage devices. Table 6 shows these new metrics and their values for DRAM, HDD, and tape. In addition to Kaps, Maps, and scan time, the table also shows $/Kaps, $/Maps, and $/TBscan, where costs are amortized over a three-year time frame as proposed by Gray and Graefe.[8]

Looking at $/Kaps, we see that DRAM is five orders of magnitude cheaper than HDD, which, in turn, is six orders of magnitude cheaper than tape. This is expected given the huge disparity in random access latencies and is in accordance with the five-minute rule that favors using DRAM for randomly accessed data. Looking at $/Maps, we see that the difference between DRAM and HDD shrinks to roughly 1,000×. This is due to the fact that HDDs can provide much higher throughput for sequential data accesses over random ones. However, HDD continue to be six orders of magnitude cheaper than tape even for MB-sized random data accesses. This, also, is in accordance with the HDD/tape asymptote shown in Figure 2. Finally, $/TBscan paints a very different picture. While DRAM remains 300× cheaper than HDD, the difference between HDD and tape shrinks to 10×.

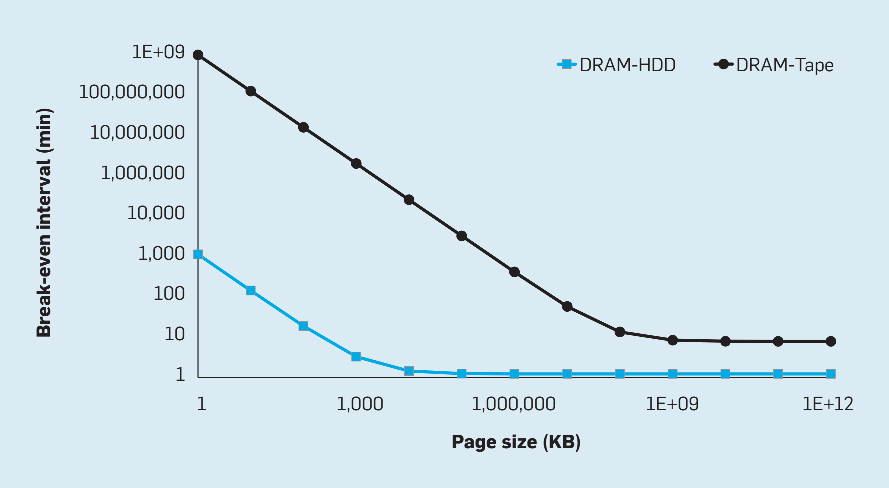
Comparing the $/TBscan values with those reported in 1997, we can see two interesting trends. First, the disparity between DRAM and HDD is growing over time. In 1997, it was 13× cheaper to use DRAM for a TBscan than HDD. Today, it is 300× cheaper. This implies that even for scan-intensive applications, unsurprisingly, optimizing for performance requires avoiding using HDD as the storage medium. Second, the difference between HDD and tape is following the opposite trend and shrinking over time. In 1997, HDD was 70× cheaper than tape. However, today it is only 10× cheaper. Unlike HDD, sequential data transfer bandwidth of tape is predicted to double for the foreseeable future. Hence, this difference is likely to shrink further. Thus, in the near future, it might not make much of a difference whether data is stored in a tape or HDD with respect to the price paid per TB scan.

Implications. Today, all data generated by an enterprise has to be stored twice, once in the traditional HDD-based capacity tier for enabling batch analytics, and a second time in the tape-based archival tier for meeting regulatory compliance requirements. The shrinking difference in $/TBscan between HDD and tape suggests that it might be economically beneficial to merge the capacity and archival tiers into a single *cold storage tier*.[3] However, with such a merger, the cold storage tier would no longer be a near-line tier that is used rarely during disaster recovery, but an online tier that is used for running batch analytics applications. Recent hardware and application trends indicate that it might be feasible to build such a cold storage tier.

On the hardware front, storage vendors have recently started building new *cold storage devices* (CSD) for storing cold data. Each CSD is an ensemble of HDDs grouped in a massive array of idle disks (MAID) setup where only a small subset of disks are active at any given time.[2,4,27] For instance, Pelican CSD pro vides 5PB of storage using 1,152 SMR disks packed as a 52U rack appliance.[2] However, only 8% of disks can be spun up simultaneously due to cooling and power restrictions enforced by hardware. Access to data in any of the spun-up disks can be done with latency and bandwidth comparable to that of the traditional capacity tier. For instance, Pelican, OpenVault Knox, and ArticBlue provide between 1–2GB/s of throughput for reading data from spun-up disks.[2,21,27] However, accessing data on a spun-down disk takes several seconds, as the disk has to be spun up before data can be retrieved. Thus, CSDs form a perfect middle ground between HDD and tape with respect to both cost/GB and access latency.

On the application front, there is a clear bifurcation in demand between latency-sensitive interactive applications and latency insensitive batch applications. As interactive applications are isolated to the performance tier, the cold storage tier only has to cater to the bandwidth demands of latency-insensitive batch analytics applications. Nearline storage devices like tape libraries and CSD are capable of providing high-throughput access for sequentially accessed data. Thus, researchers have recently started investigating extensions to batch processing frameworks for enabling analytics directly over data stored in tape archives and CSD. For instance, Nakshatra implements prefetching and I/O scheduling extensions to Hadoop so that mapreduce jobs can be scheduled to run directly on tape archives.[14] Skipper is a query-processing framework that uses adaptive query processing techniques in combination with customized caching and I/O scheduling to enable que-

Figure 2. Break-even interval asymptotes for DRAM–HDD and DRAM–tape cases.



ry execution over CSD.[3] Skipper even shows that for long-running batch queries, using CSD results in query execution time increasing by only 35% compared to a traditional HDD despite the long disk spin-up latency. With such frameworks, it should be possible for installations to switch from the traditional three-tier hierarchy to a two-tier hierarchy consisting of just a performance tier with DRAM and SSDs, and a cold storage tier with CSDs.

Conclusion and Future Work

Modern database engines use a three-tier storage hierarchy across four primary storage media (DRAM, SSD, HDD, and tape) with widely varying price-performance characteristics. In this article, we revisited the five-minute rule in the context of this modern storage hierarchy and used it to highlight impending changes based on recent trends in the hardware landscape.

In the performance tier, NAND flash is inching its way closer to the CPU resulting in dramatic improvements in both access latency and bandwidth. For state-of-the-art PCIe SSDs, the break-even interval predicted by the five-minute rule is one minute for 4KB pages. Going forward, further improvements in NAND flash and the introduction of new NVM technologies will likely result in this interval dropping further. As the data reuse window shrinks, it will soon be economically more valuable to store most, if not all, data on solid-state storage devices instead of DRAM. This will invariably necessitate revisiting several techniques pioneered by traditional HDD-based database engines, but eschewed by in-memory engines, like buffer caching, on-disk storage layout, and index persistence, to name a few, for these new low-latency, high-bandwidth storage devices.

Traditionally, HDDs have been used for implementing the capacity tier. However, our analysis showed that the difference between HDD and tape is shrinking when $/TBScan is used as the metric. Given the latency-insensitive nature of batch analytics workloads, it is economically beneficial to merge the HDD-based capacity tier and the tape-based archival tier into a single cold storage tier as demonstrated by recent research.[3] However, several open questions still need to answered in order for the cold storage tier to be feasible in practice.

Over the past few years, several other systems have been built to reduce the cost of storing cold data using alternative storage media. For instance, DT-Store[16] uses LTFS tape archive for reducing the TCO of online multimedia streaming services by storing cold data in tape drives. ROS[28] is a PB-sized, rack-scale cold storage library built using thousands of optical discs packed in a single 42U Rack. Today, it is unclear as to how these alternative storage options fare with respect to HDD-based CSD as the storage media of choice for storing cold data. Furthermore, in order for the Cold Storage Tier to be realized in practice, an ideal cold storage media needs to support batch analytics workloads. CSD, tape, and optical media are all primarily used today for archival storage where data is rarely read. Further research is required to understand the reliability implications of using these storage devices under batch analytics workloads.

Finally, with widespread adoption of cloud computing, the modern enterprise storage hierarchy not only spans several storage devices, but also different geographic locations from direct-attached low-latency devices, through network-attached storage servers, to cloud-hosted storage services. The price-performance characteristics of these storage configurations vary dramatically depending not only on the storage media used, but also on other factors like the total capacity of data stored, the frequency and granularity of I/O operations used to access the data, the read–write ratio, the duration of data storage, and the cloud service provider used, to name a few. Given the multitude of factors, determining the break-even interval for cloud storage is a complicated problem that we did not consider in this work. Thus, another interesting avenue of future work is extending the five-minute rule to such a distributed cloud storage setting. C

References

1. Appuswamy, R., Olma, M., and Ailamaki, A. Scaling the memory power wall with dram-aware data management. In *Proceedings of DaMoN*, 2015.
2. Balakrishnan, S. et al. Pelican: A building block for exascale cold data storage. In *Proceedings of OSDI*, 2014.
3. Borovica-Gajic, R., Appuswamy, R., and Ailamaki, A. Cheap data analytics using cold storage devices. In *Proceedings of VLDB 9*, 12 (2016).
4. Colarelli, D. and Grunwald, D. Massive arrays of idle disks for storage archives. In *Proceedings of 2002 Conference on Supercomputing*.
5. Coughlin, T. Flash memory areal densities exceed those of hard drives; http://bit.ly/2NbDh5T.
6. Graefe, G. The five-minute rule 20 years later (and how flash memory changes the rules). *Commun. ACM 52*, 7 (July 2009).
7. Gray, J. The five-minute rule; research.microsoft.com/en-us/um/people/gray/talks/fiveminuterule.ppt.
8. Gray, J. and Graefe, G. The five-minute rule ten years later, and other computer storage rules of thumb. *SIGMOD Rec. 26*, 4 (1997).
9. Gray, J. and Putzolu, F. The 5-minute rule for trading memory for disc accesses and the 10-byte rule for trading memory for CPU time. In *Proceedings of SIGMOD*, 1987.
10. Horison Information Strategies Report. Tiered storage takes center stage, IDC. Technology assessment: Cold storage is hot again—finding the frost point; http://www.storiant.com/resources/Cold-Storage-Is-Hot-Again.pdf.
11. Intel. Cold Storage in the Cloud: Trends, Challenges, and Solutions, 2013; https://intel.ly/2ZG74F6.
12. I.S.I. Consortium. International magnetic tape storage roadmap; http://www.insic.org/news/2015roadmap/15index.html
13. Kathpal, A. and Yasa, G.A.N. Nakshatra: Towards running batch analytics on an archive. In *Proceedings of MASCOTS*, 2014.
14. Lantz, M. Why the future of data storage is (still) magnetic tape; http://bit.ly/2XChrMO
15. Lee, J., Ahn, J., Park, C., and Kim, J. Dtstorage: Dynamic tape-based storage for cost-effective and highly-available streaming service. In *Proceedings of CCGRID*, 2016.
16. Lim, K., Chang, J., Mudge, T., Ranganathan, P., Reinhardt, S.K., and Wenisch, T.F. Disaggregated memory for expansion and sharing in blade servers. In *Proceedings of ISCA*, 2009.
17. Lomet, D. Cost/performance in modern data stores: How data caching systems succeed. In *Proceedings of DaMoN*, 2018.
18. Moore, F. Storage outlook 2016; http://bit.ly/2KBLgao.
19. Perlmutter, M. The lost picture show: Hollywood archivists cannot outpace obsolescence, 2017; http://bit.ly/2KDaqWd.
20. Spectra. Arcticblue deep storage disk. Product, https://www.spectralogic.com/products/arcticblue/.
21. SpectraLogic. Spectralogic t50e; http://bit.ly/2Ych8pl.
22. StorageReview. Intel optane memory review. http://www.storagereview.com/intel_optane_memory_review.
23. TPC-C. Dell-microsoft sql server tpc-c executive summary, 2014; http://www.tpc.org/tpcc/results/tpcc_result_ detail.asp?id=114112501.
24. Ultrium. LTP ultrium roadmap;
25. http://www.ltoultrium.com/lto-ultrium-roadmap/.
26. Umamageswaran, K. and Goindi, G. Exadata: Delivering memory performance with shared flash; http://bit.ly/2LhBVEa.
27. Yan, M. Open compute project: Cold storage hardware v0.5, 2013; http://bit.ly/2X6H2Ot.
28. Yan, W., Yao, J., Cao, Q., Xie, C., and Jiang, H. Ros: A rack-based optical storage system with inline accessibility for long-term data preservation. In *Proceedings of EUROSYS*, 2017.

Raja Appuswamy (raja.appuswamy@eurecom.fr) is an assistant professor in the Data Science Department at EURECOM, Biot, Provence-Alpes-Côte d'Azur, France.

Goetz Graefe (goetzg@google.com), Google, Inc., Madison, WI, USA.

Renata Borovica-Gajic (renata.borovica@unimelb.edu.au) is an assistant professor in the School of Computing and Information Systems at the University of Melbourne, Australia.

Anastasia Ailamaki (anastasia.ailamaki@epfl.ch) is a professor at EPFL, Lausanne, Switzerland, and director of its Data-Intensive Applications and Systems (DIAS) lab.

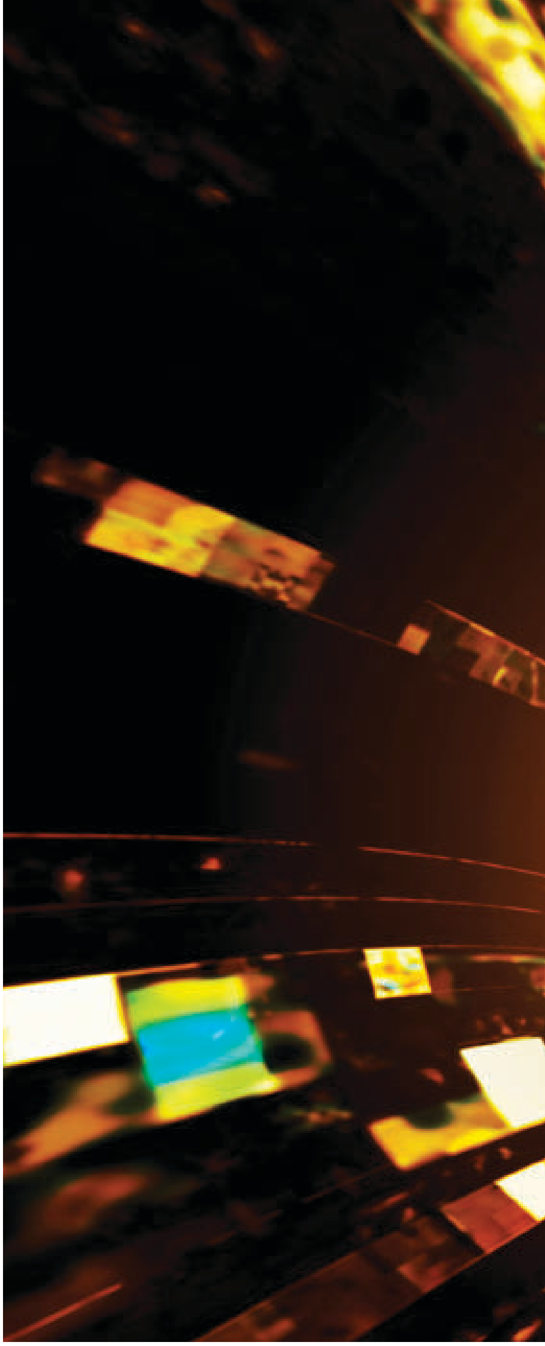
Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers
Bryan Jeffrey Parno
Text Data Management and Analysis
A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai
Sean Massung
The Sparse Fourier Transform
Theory and Practice
Haitham Hassanieh
Frontiers of Multimedia Research
Shih-Fu Chang
Editor
Smarter than Their Machines
Oral Histories of Pioneers in Interactive Computing
John Cullinane
The VR Book
Human-Centered Design for Virtual Reality
Jason Jerald, Ph.D.
Verified Functional Programming in Agda
Aaron Stump
A Framework for Scientific Discovery through Video Games
Seth Cooper
An Architecture for Fast and General Data Processing on Large Clusters
Matei Zaharia
The Handbook of Multimodal-Multisensor Interfaces
Volume 1
Foundations, User Modeling, and Common Modality Combinations

DOI:10.1145/3363294

Demystifying the uses of a powerful tool for uncertain information.

BY YAN PEI, SWARNENDU BISWAS, DONALD S. FUSSELL, AND KESHAV PINGALI

An Elementary Introduction to Kalman Filtering



KALMAN FILTERING IS a state estimation technique used in many application areas such as spacecraft navigation, motion planning in robotics, signal processing, and wireless sensor networks because of its ability to extract useful information from noisy data and its small computational and memory requirements.[12,20,27–29] Recent work has used Kalman filtering in controllers for computer systems.[5,13,14,23]

Although many introductions to Kalman filtering are available in the literature,[1–4,6–11,17,21,25,29] they are usually focused on particular applications such as robot motion or state estimation in linear systems, making it difficult to see how to apply Kalman filtering to other problems. Other presentations derive Kalman filtering as an application of Bayesian inference, assuming that noise is Gaussian. This leads to the common misconception that Kalman filtering can be applied only if noise is Gaussian.[15]

Abstractly, Kalman filtering can be seen as a particular approach to combining approximations of an unknown value to produce a better approximation. Suppose we use two devices of different

» key insights

- This article presents an elementary derivation of Kalman filtering, a classic state estimation technique.
- Understanding Kalman filtering is useful for more principled control of computer systems.
- Kalman filtering is used as a black box by many computer scientists.

IMAGE BY A LUNA BLUE

designs to measure the temperature of a CPU core. Because devices are usually noisy, the measurements are likely to differ from the actual temperature of the core. As the devices are of different designs, let us assume that noise affects the two devices in unrelated ways (this is formalized here using the notion of correlation). Therefore, the measurements x_1 and x_2 are likely to be different from each other and from the actual core temperature x_c. A natural question is the following: is there a way to combine the information in the noisy measurements x_1 and x_2 to obtain a good approximation of the actual temperature x_c?

One ad hoc solution is to use the formula $0.5{*}x_1{+}0.5{*}x_2$ to take the average of the two measurements, giving them equal weight. Formulas of this sort are called *linear estimators* because they use a weighted sum to fuse values; for our temperature problem, their general form is $\beta{*}x_1{+}\alpha{*}x_2$. In this presentation, we use the term *estimate* to refer to both a noisy measurement and a value computed by an estimator, as both are approximations of unknown values of interest.

Suppose we have additional information about the two devices, say the second one uses more advanced temperature sensing. Because we would have more confidence in the second measurement, it seems reasonable that we should discard the first one, which is equivalent to using the linear estimator $0.0{*}x_1 + 1.0{*}x_2$. Kalman filtering tells us that in general, this intuitively reasonable linear estimator is not "optimal;" paradoxically, there is useful information even in the measurement from the lower quality device, and the optimal estimator is one in which the weight given to each measurement is proportional to the confidence we have in the device producing that measurement. Only if we have no confidence whatever in the first device should we discard its measurement.

The goal of this article[a] is to present the abstract concepts behind Kalman filtering in a way that is accessible to most computer scientists while clarifying the key assumptions, and then show how the problem of state estimation in linear systems can be solved as an

[a] An extended version of this article that includes additional background material and proofs is available.[30]

application of these general concepts. First, the informal ideas discussed here are formalized using the notions of distributions and random samples from distributions. Confidence in estimates is quantified using the variances and covariances of these distributions.[b] Two algorithms are described next. The first one shows how to fuse estimates (such as core temperature measurements) optimally, given a reasonable definition of optimality. The second algorithm addresses a problem that arises frequently in practice: estimates are vectors (for example, the position and velocity of a robot), but only a part of the vector can be measured directly; in such a situation, how can an estimate of the entire vector be obtained from an estimate of just a part of that vector? The best linear unbiased estimator (BLUE) is used to solve this problem.[16,19,26] It is shown that the Kalman filter can be derived in a straightforward way by using these two algorithms to solve the problem of state estimation in linear systems. The extended Kalman filter and unscented Kalman filter, which extended Kalman filtering to nonlinear systems, are described briefly at the end of the article.

Formalizing Estimates

Scalar estimates. To model the behavior of devices producing noisy temperature measurements, we associate each device i with a *random variable* that has a *probability density function* (pdf) $p_i(x)$ such as the ones shown in Figure 1 (the x-axis in this figure represents temperature). Random variables need not be Gaussian.[c] Obtaining a measurement from device i corresponds to drawing a random sample from the distribution for that device. We write $x_1 \sim p_1(\mu_1, \sigma_1^2)$ to denote that x_i is a random variable with pdf p_i whose mean and variance are μ_i and σ_i^2, respectively; following convention, we use x_i to represent a random sample from this distribution as well.

Means and variances of distributions model different kinds of inaccuracies in measurements. Device i is said to have a *systematic error* or *bias* in its measurements if the mean μ_i of its distribution is not equal to the actual temperature x_c (in general, to the value being estimated, which is known as *ground truth*); otherwise, the instrument is unbiased. Figure 1 shows pdfs for two devices that have different amounts of systematic error. The variance σ_i^2 on the other hand is a measure of the *random error* in the measurements. The impact of random errors can be mitigated by taking many measurements with a given device and averaging their values, but this approach will not reduce systematic error.

In the formulation of Kalman filtering, it is assumed that measuring devices do not have systematic errors. However, we do not have the luxury of taking many measurements of a given state, so we must take into account the impact of random error on a single measurement. Therefore, confidence in a device is modeled formally by the variance of the distribution associated with that device; the smaller the variance, the higher our confidence in the measurements made by the device. In Figure 1, the fact we have less confidence in the first device has been illustrated by making p_1 more spread out than p_2, giving it a larger variance.

The informal notion that noise should affect the two devices in "unrelated ways" is formalized by requiring that the corresponding random variables be *uncorrelated*. This is a weaker condition than requiring them to be *independent*, as explained in our online appendix (http://dl.acm.org/citation.cfm?doid=3363294&picked=formats). Suppose we are given the measurement made by one of the devices (say x_1) and we have to guess what the other measurement (x_2) might be. If knowing x_1 does not give us any new information about what x_2 might be, the random variables are independent. This is expressed formally by the equation $p(x_2|x_1) = p(x_2)$; intuitively, knowing the value of x_1 does not change the pdf for the possible values of x_2. If the random variables are only uncorrelated, knowing x_1 might give us new information about x_2 such as restricting its possible values but the mean of $x_2|x_1$ will still be μ_2. Using expectations, this can be written as $E[x_2|x_1] = E[x_2]$, which is equivalent to requiring that $E[(x_1 - \mu_1)(x_2 - \mu_2)]$, the covariance between the two variables, be equal to zero. This is obviously a weaker condition than independence.

Although the discussion in this section has focused on measurements, the same formalization can be used for estimates produced by an estimator. Lemma 1(i) shows how the mean and variance of a linear combination of pairwise uncorrelated random variables can be computed from the means and variances of the random variables.[18] The mean and variance can be used to quantify bias and random errors for the estimator as in the case of measurements.

An *unbiased estimator* is one whose mean is equal to the unknown value being estimated and it is preferable to a biased estimator with the same variance. Only unbiased estimators are considered in this article. Furthermore, an unbiased estimator with a smaller variance is preferable to one with a larger variance as we would have more confidence in the estimates it produces. As a step toward generalizing this discussion to estimators that produce vector estimates, we refer to the variance of an unbiased scalar estimator as the *mean square error* of that estimator or *MSE* for short.

Lemma 1(ii) asserts that if a random variable is pairwise uncorrelated with a set of random variables, it is uncorrelated with any linear combination of those variables.

Lemma 1. *Let $x_1 \sim p_1(\mu_1, \sigma_1^2), .., x_n \sim p_n(\mu_n, \sigma_n^2)$ be a set of pairwise uncorrelated random variables. Let $y = \sum_{i=1}^{n} \alpha_i x_i$ be a random variable that is a linear combination of the x_i's.*

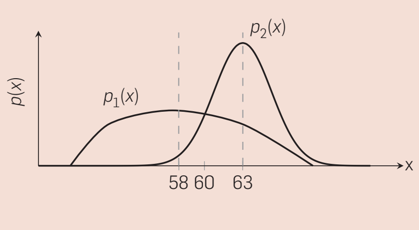
(i) The mean and variance of y are:

$$\mu_y = \sum_{i=1}^{n} \alpha_i \mu_i \qquad (1)$$

$$\sigma_y^2 = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 \qquad (2)$$

(ii) If random variable x_{n+1} is pair-wise uncorrelated with $x_1, .., x_n$, it is uncorrelated with y.

Figure 1. Using pdfs to model devices with systematic and random errors. Ground truth is 60°C. Dashed lines are means of pdfs.



[b] Basic concepts such as probability density function, mean, expectation, variance and covariance are introduced in the online appendix.

[c] The role of Gaussians in Kalman filtering is discussed later in the article.

Vector estimates. In some applications, estimates are vectors. For example, the state of a mobile robot might be represented by a vector containing its position and velocity. Similarly, the vital signs of a person might be represented by a vector containing his temperature, pulse rate, and blood pressure. Here, we denote a vector by a boldfaced lowercase letter, and a matrix by an uppercase letter.

The covariance matrix $\sum_{\mathbf{xx}}$ of a random variable $\mathbf{x}$ is the matrix $E[(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^{\mathrm{T}}]$, where $\boldsymbol{\mu}_{\mathbf{x}}$ is the mean of $\mathbf{x}$. Intuitively, entry (i,j) of this matrix is the covariance between the i and j components of vector $\mathbf{x}$; in particular, entry (i,i) is the variance of the i^{th} component of $\mathbf{x}$. A random variable $\mathbf{x}$ with a pdf p whose mean is $\boldsymbol{\mu}_x$ and covariance matrix is $\sum_{\mathbf{xx}}$ is written as $\mathbf{x}{\sim}p(\boldsymbol{\mu}_x, \sum_{\mathbf{xx}})$. The inverse of the covariance matrix $(\Sigma_{\mathbf{xx}}^{-1})$ is called the precision or *information* matrix.

Uncorrelated random variables. The cross-covariance matrix $\sum_{\mathbf{vw}}$ of two random variables $\mathbf{v}$ and $\mathbf{w}$ is the matrix $E[(\mathbf{v}{-}\boldsymbol{\mu}_{\mathbf{v}})(\mathbf{w}{-}\boldsymbol{\mu}_{w})^{\mathrm{T}}]$. Intuitively, element (i,j) of this matrix is the covariance between elements $\mathbf{v}(i)$ and $\mathbf{w}(j)$. If the random variables are uncorrelated, all entries in this matrix are zero, which is equivalent to saying that every component of $\mathbf{v}$ is uncorrelated with every component of $\mathbf{w}$. Lemma 2 generalizes Lemma 1.

Lemma 2. *Let* $\mathbf{x}_1{\sim}p_1(\boldsymbol{\mu}_1, \sum_1)$, …, $\mathbf{x}_n{\sim}p_n(\boldsymbol{\mu}_n, \sum_n)$ *be a set of pairwise uncorrelated random variables of length m. Let* $\boldsymbol{y} = \sum_{i=1}^{n} A_i\,\boldsymbol{x}_i$.

(i) The mean and covariance matrix of $\boldsymbol{y}$ *are the following:*

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \sum_{i=1}^{n} A_i\,\boldsymbol{\mu}_i \qquad (3)$$

$$\Sigma_{\boldsymbol{yy}} = \sum_{i=1}^{n} A_i\,\Sigma_i\,A_i^{\mathrm{T}} \qquad (4)$$

(ii) If random variable $\boldsymbol{x}_{n+1}$ *is pairwise uncorrelated with* $\boldsymbol{x}_1$, .., $\boldsymbol{x}_n$, *it is uncorrelated with* $\mathbf{y}$.

The *MSE* of an unbiased estimator $\mathbf{y}$ is $E[(\mathbf{y}{-}\boldsymbol{\mu}_{\mathbf{y}})^{\mathrm{T}}(\mathbf{y}{-}\boldsymbol{\mu}_{\mathbf{y}})]$, which is the sum of the variances of the components of $\mathbf{y}$; if $\mathbf{y}$ has length 1, this reduces to variance as expected. The *MSE* is also the sum of the diagonal elements of $\sum_{\mathbf{yy}}$ (this is called the *trace* of $\sum_{\mathbf{yy}}$).

An unbiased estimator is one whose mean is equal to the unknown value being estimated and it is preferable to a biased estimator with the same variance.

Fusing Scalar Estimates

We now consider the problem of choosing the optimal values of the parameters α and β in the linear estimator $\beta{*}x_1 + \alpha{*}x_2$ for fusing two estimates x_1 and x_2 from uncorrelated scalar-valued random variables.

The first reasonable requirement is that if the two estimates x_1 and x_2 are equal, fusing them should produce the same value. This implies that $\alpha{+}\beta$ =1. Therefore, the linear estimators of interest are of the form

$$y_\alpha(x_1,x_2) = (1-\alpha)*x_1 + \alpha*x_2 \qquad (5)$$

If x_1 and x_2 in Equation 5 are considered to be unbiased estimators of some quantity of interest, then y_α is an unbiased estimator for any value of α. How should optimality of such an estimator be defined? One reasonable definition is that the optimal value of α *minimizes the variance of* y_α as this will produce the highest-confidence fused estimates.

Theorem 1. *Let* $x_1{\sim}p_1(\mu_1,\sigma_1^2)$ *and* $x_2{\sim}p_2(\mu_2,\sigma_2^2)$ *be uncorrelated random variables. Consider the linear estimator* $y_\alpha(x_1,x_2){=}(1{-}\alpha){*}x_1{+}\alpha{*}x_2$. *The variance of the estimator is minimized for* $\alpha = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}$.

The proof is straightforward and is given in the online appendix. The variance (*MSE*) of y_α can be determined from Lemma 1:

$$\sigma_y^2(\alpha) = (1-\alpha)^2 * \sigma_1^2 + \alpha^2 * \sigma_2^2 \qquad (6)$$

Setting the derivative of $\sigma_y^2(\alpha)$ with respect to α to zero and solving the resulting equation yield the required result.

In the literature, the optimal value of α is called the *Kalman gain K*. Substituting K into the linear fusion model, we get the optimal linear estimator $y(x_1, x_2)$:

$$y(x_1,x_2) = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} * x_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} * x_2 \qquad (7)$$

As a step toward fusion of n>2 estimates, it is useful to rewrite this as follows:

$$y(x_1,x_2) = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}} * x_1 + \frac{\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}} * x_2 \qquad (8)$$

Substituting the optimal value of α into Equation 6, we get

$$\sigma_y^2 = \frac{1}{\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}} \qquad (9)$$

The expressions for y and σ_y^2 are complicated because they contain the reciprocals of variances. If we let ν_1 and ν_2 denote the precisions of the two distributions, the expressions for y and ν_y can be written more simply as follows:

$$y(x_1,x_2) = \frac{\nu_1}{\nu_1+\nu_2} * x_1 + \frac{\nu_2}{\nu_1+\nu_2} * x_2 \qquad (10)$$

$$\nu_y = \nu_1 + \nu_2 \qquad (11)$$

These results say the weight we should give to an estimate is proportional to the confidence we have in that estimate, and that we have more confidence in the fused estimate than in the individual estimates, which is intuitively reasonable. To use these results, we need only the variances of the distributions. In particular, the pdfs p_i, which are usually not available in applications, are not needed, and the proof of Theorem 1 does not require these pdfs to have the same mean.

Fusing multiple scalar estimates. These results can be generalized to optimally fuse multiple pairwise uncorrelated estimates x_1, x_2, ..., x_n. Let $y_{n,\alpha}(x_1, .., x_n)$ denote the linear estimator for fusing the n estimates given parameters α_1, .., α_n, which we denote by α (the notation $y_\alpha(x_1, x_2)$ introduced previously can be considered to be an abbreviation of $y_{2,\alpha}(x_1, x_2)$).

Theorem 2. *Let $x_i \sim p_i(\mu_i,\sigma_i^2)$ for $(1 \le i \le n)$ be a set of pairwise uncorrelated random variables. Consider the linear estimator $y_{n,\alpha}(x_1,..,x_n)=\sum_{i=1}^{n}\alpha_i x_i$ where $\sum_{i=1}^{n}\alpha_i = 1$. The variance of the estimator is minimized for*

$$\alpha_i = \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{n}\frac{1}{\sigma_j^2}}$$

The minimal variance is given by the following expression:

$$\sigma_{y_n}^2 = \frac{1}{\sum_{j=1}^{n}\frac{1}{\sigma_j^2}} \qquad (12)$$

As before, these expressions are more intuitive if the variance is replaced with precision: the contribution of x_i to the value of $y_n(x_1, .., x_n)$ is proportional to x_i's confidence.

Kalman filtering can be seen as a particular approach to combining approximations of an unknown value to produce a better approximation.

$$y_n(x_1,..,x_n) = \sum_{i=1}^{n}\frac{\nu_i}{\nu_1+\ldots+\nu_n} * x_i \qquad (13)$$

$$\nu_{y_n} = \sum_{i=1}^{n}\nu_i \qquad (14)$$

Equations 13 and 14 generalize Equations 10 and 11.

Incremental fusing is optimal. In many applications, the estimates x_1, x_2, ..., x_n become available successively over a period of time. Although it is possible to store all the estimates and use Equations 13 and 14 to fuse all the estimates from scratch whenever a new estimate becomes available, it is possible to save both time and storage if one can do this fusion incrementally. We show that just as a sequence of numbers can be added by keeping a running sum and adding the numbers to this running sum one at a time, a sequence of n>2 estimates can be fused by keeping a "running estimate" and fusing estimates from the sequence one at a time into this running estimate without any loss in the quality of the final estimate. In short, we want to show that $y_n(x_1,..,x_n)=y_2(y_2(..y_2(x_1,x_2)...),x_n)$. A little bit of algebra shows that if n>2, Equations 13 and 14 for the optimal linear estimator and its precision can be expressed as shown in Equations 15 and 16.

$$y_n(x_1,..,x_n) = \frac{\nu_n}{\nu_{y_{n-1}}+\nu_n}x_n + \frac{\nu_{y_{n-1}}}{\nu_{y_{n-1}}+\nu_n}y_{n-1}(x_1,\ldots,x_{n-1}) \qquad (15)$$

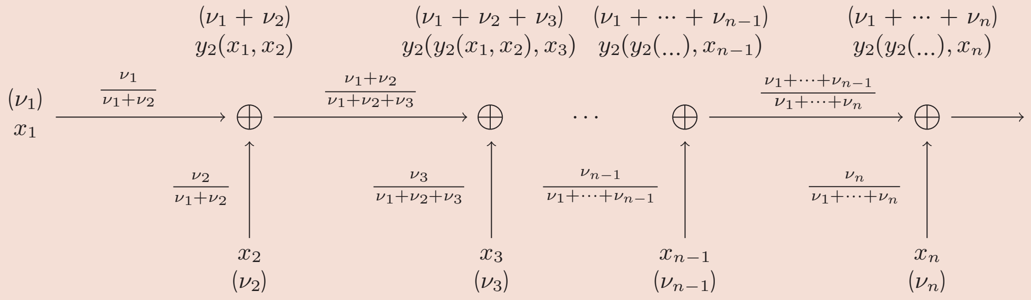
$$\nu_{y_n} = \nu_{y_{n-1}} + \nu_n \qquad (16)$$

This shows that $y_n(x_1,..,x_n) = y_2(y_{n-1}(x_1,..,x_{n-1}), x_n)$. Using this argument recursively gives the required result.[d]

To make the connection to Kalman filtering, it is useful to derive the same result using a pictorial argument. Figure 2 shows the process of incrementally fusing the n estimates. In this picture, time progresses from left to right, the precision of each estimate is shown in parentheses next to it, and the weights on the edges are the weights from Equation 10. The contribution of each x_i to the final value $y_2(y_2(\ldots), x_n)$ is given by the product of the weights on the path from x_i to the final value, and this product is obviously equal to the weight of x_i in

[d] We thank Mani Chandy for showing us this approach to proving the result.

Figure 2. Dataflow graph for incremental fusion.


(ν_1)
x_1
$\frac{\nu_1}{\nu_1+\nu_2}$
$\frac{\nu_2}{\nu_1+\nu_2}$
x_2
(ν_2)
$(\nu_1+\nu_2)$
$y_2(x_1,x_2)$
$\frac{\nu_1+\nu_2}{\nu_1+\nu_2+\nu_3}$
$\frac{\nu_3}{\nu_1+\nu_2+\nu_3}$
x_3
(ν_3)
$(\nu_1+\nu_2+\nu_3)$
$y_2(y_2(x_1,x_2),x_3)$
$\cdots$
$\frac{\nu_{n-1}}{\nu_1+\cdots+\nu_{n-1}}$
x_{n-1}
(ν_{n-1})
$(\nu_1+\cdots+\nu_{n-1})$
$y_2(y_2(\ldots),x_{n-1})$
$\frac{\nu_1+\cdots+\nu_{n-1}}{\nu_1+\cdots+\nu_n}$
$\frac{\nu_n}{\nu_1+\cdots+\nu_n}$
x_n
(ν_n)
$(\nu_1+\cdots+\nu_n)$
$y_2(y_2(\ldots),x_n)$

Equation 13, showing that incremental fusion is optimal.

Summary. The results in this section can be summarized informally as follows. *When using a linear estimator to fuse uncertain scalar estimates, the weight given to each estimate should be inversely proportional to the variance of the random variable from which that estimate is obtained. Furthermore, when fusing $n>2$ estimates, estimates can be fused incrementally without any loss in the quality of the final result.* These results are often expressed formally in terms of the Kalman gain K, as shown in Figure 3; the equations can be applied recursively to fuse multiple estimates. Note that if $\nu_1 \gg \nu_2$, $K \approx 0$ and $y(x_1,x_2) \approx x_1$; conversely if $\nu_1 \ll \nu_2$, $K \approx 1$ and $y(x_1,x_2) \approx x_2$.

Fusing Vector Estimates

The results for fusing scalar estimates can be extended to vectors by replacing *variances* with *covariance matrices*.

For vectors, the linear estimator is $\mathbf{y}_{n,A}(\mathbf{x}_1,\mathbf{x}_2,..,\mathbf{x}_n)=\sum_{i=1}^{n} A_i\mathbf{x}_i$ where $\sum_{i=1}^{n} A_i = I$. Here A stands for the matrix parameters $(A_1,\ldots,A_n)$. All the vectors $(\mathbf{x}_i)$ are assumed to be of the same length. To simplify notation, we omit the subscript n in $\mathbf{y}_{n,A}$ in the discussion here as it is obvious from the context.

Optimality. The parameters A_1, ..., A_n in the linear data fusion model are chosen to minimize $MSE(\mathbf{y}_A)$ which is $E[(\mathbf{y}_A-\boldsymbol{\mu}_{yA})^{\mathrm{T}}(\mathbf{y}_A-\boldsymbol{\mu}_{yA})]$.

Theorem 3 generalizes Theorem 2 to the vector case. The proof of this theorem is given in the appendix. Comparing Theorems 2 and 3, we see that the expressions are similar, the main difference being that the role of variance in the scalar case is played by the covariance matrix in the vector case.

Theorem 3. *Let $\mathbf{x}_i \sim p_i(\boldsymbol{\mu}_i, \Sigma_i)$ for $(1 \le i \le n)$ be a set of pairwise uncorrelated random variables. Consider the linear estimator $\mathbf{y}_A(\mathbf{x}_1,..,\mathbf{x}_n)=\sum_{i=1}^{n} A_i\mathbf{x}_i$, where $\sum_{i=1}^{n} A_i = I$. The value of $MSE(\mathbf{y}_A)$ is minimized for*

$$A_i = \Big(\sum_{j=1}^{n}\Sigma_j^{-1}\Big)^{-1}\Sigma_i^{-1} \qquad (23)$$

Therefore the optimal estimator is

$$\mathbf{y}(\mathbf{x}_1,\ldots,\mathbf{x}_n) = \Big(\sum_{j=1}^{n}\Sigma_j^{-1}\Big)^{-1}\sum_{i=1}^{n}\Sigma_i^{-1}\mathbf{x}_i \qquad (24)$$

The covariance matrix of $\mathbf{y}$ can be computed by using Lemma 2.

$$\Sigma_{\mathbf{yy}} = \Big(\sum_{j=1}^{n}\Sigma_j^{-1}\Big)^{-1} \qquad (25)$$

In the vector case, precision is the inverse of a covariance matrix, denoted by N. Equations 26–27 use precision to express the optimal estimator and its variance and generalize Equations 13–14 to the vector case.

$$\mathbf{y}(\mathbf{x}_1,\ldots,\mathbf{x}_n) = N_{\mathbf{y}}^{-1}\sum_{i=1}^{n} N_i\mathbf{x}_i \qquad (26)$$

$$N_{\mathbf{y}} = \sum_{j=1}^{n} N_j \qquad (27)$$

As in the scalar case, fusion of $n>2$ vector estimates can be done incrementally without loss of precision. The proof is similar to the scalar case and is omitted.

Figure 3. Optimal fusion of scalar estimates.

$$x_1 \sim p_1(\mu_1,\sigma_1^2), \quad x_2 \sim p_2(\mu_2,\sigma_2^2)$$

$$K = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} = \frac{\nu_2}{\nu_1+\nu_2} \qquad (17)$$

$$y(x_1,x_2) = x_1 + K(x_2 - x_1) \qquad (18)$$

$$\sigma_y^2 = (1-K)\sigma_1^2 \quad \text{or} \quad \nu_y = \nu_1+\nu_2 \qquad (19)$$

Figure 4. Optimal fusion of vector estimates.

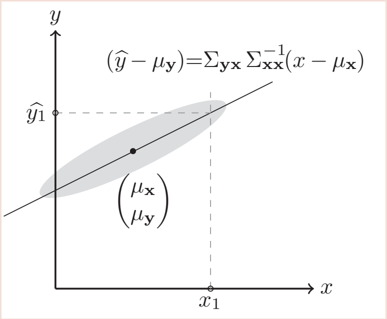
$$\mathbf{x}_1 \sim p_1(\boldsymbol{\mu}_1,\Sigma_1), \quad \mathbf{x}_2 \sim p_2(\boldsymbol{\mu}_2,\Sigma_2)$$

$$K = \Sigma_1(\Sigma_1+\Sigma_2)^{-1} = (N_1+N_2)^{-1}N_2 \qquad (20)$$

$$\mathbf{y}(\mathbf{x}_1,\mathbf{x}_2) = \mathbf{x}_1 + K(\mathbf{x}_2 - \mathbf{x}_1) \qquad (21)$$

$$\Sigma_{\mathbf{yy}} = (I-K)\Sigma_1 \quad \text{or} \quad N_{\mathbf{y}} = N_1+N_2 \qquad (22)$$

Figure 5. BLUE line corresponding to Equation (31).



There are several equivalent expressions for the Kalman gain for the fusion of two estimates. The following one, which is easily derived from Equation 23, is the vector analog of Equation 17:

$$K = \Sigma_1(\Sigma_1 + \Sigma_2)^{-1} \qquad (28)$$

The covariance matrix of the optimal estimator $\mathbf{y}(\mathbf{x}_1, \mathbf{x}_2)$ is the following.

$$\Sigma_{yy} = \Sigma_1(\Sigma_1 + \Sigma_2)^{-1}\Sigma_2 \qquad (29)$$

$$= K\Sigma_2 = \Sigma_1 - K\Sigma_1 \qquad (30)$$

Summary. The results in this section can be summarized in terms of the Kalman gain K as shown in Figure 4.

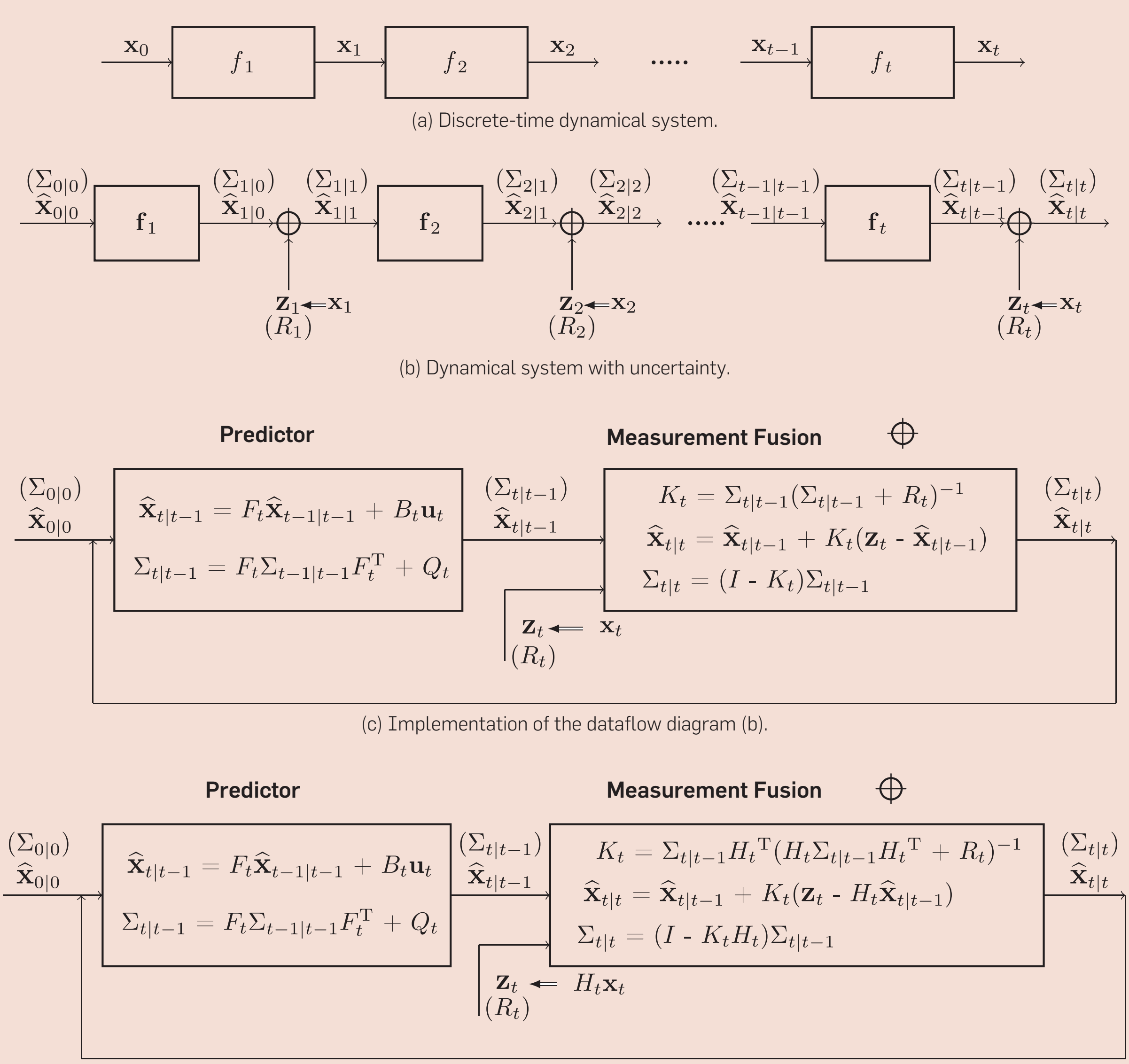
Best Linear Unbiased Estimator (BLUE)

In some applications, the state of the system is represented by a vector but only part of the state can be measured directly, so it is necessary to estimate the hidden portion of the state corresponding to a measured value of the visible state. This section describes an estimator called the *best linear unbiased estimator* (BLUE)[16,19,26] for doing this.

Consider the general problem of determining a value for vector **y** given a value for a vector **x**. If there is a functional relationship between **x** and **y** (say **y**=*F*(**x**) and *F* is given), it is easy to compute **y** given a value for **x** (say $\mathbf{x}_1$).

In our context, however, **x** and **y** are random variables, so such a precise functional relationship will not hold. Figure 5 shows an example in which *x* and *y* are scalar-valued random variables. The gray ellipse in this figure, called a *confidence ellipse*, is a projection of the joint distribution of *x* and *y* onto the (*x*, *y*) plane that shows where some large proportion of the (*x*, *y*) values are likely to be. Suppose *x* takes the value x_1. Even within the confidence ellipse, there are many points (x_1, *y*), so we cannot associate a single value of *y* with x_1. One possibility is to compute the mean of the *y* values associated with x_1 (that is,

Figure 6. State estimation using Kalman filtering.



$\mathbf{x}_0$
f_1
$\mathbf{x}_1$
f_2
$\mathbf{x}_2$
.....
$\mathbf{x}_{t-1}$
f_t
$\mathbf{x}_t$
$(\Sigma_{0|0})$ $\widehat{\mathbf{x}}_{0|0}$
$\mathbf{f}_1$
$(\Sigma_{1|0})$ $\widehat{\mathbf{x}}_{1|0}$
$(\Sigma_{1|1})$ $\widehat{\mathbf{x}}_{1|1}$
$\mathbf{f}_2$
$(\Sigma_{2|1})$ $\widehat{\mathbf{x}}_{2|1}$
$(\Sigma_{2|2})$ $\widehat{\mathbf{x}}_{2|2}$
$(\Sigma_{t-1|t-1})$ $\widehat{\mathbf{x}}_{t-1|t-1}$
$\mathbf{f}_t$
$(\Sigma_{t|t-1})$ $\widehat{\mathbf{x}}_{t|t-1}$
$(\Sigma_{t|t})$ $\widehat{\mathbf{x}}_{t|t}$
$\mathbf{z}_1 \Leftarrow \mathbf{x}_1$ (R_1)
$\mathbf{z}_2 \Leftarrow \mathbf{x}_2$ (R_2)
$\mathbf{z}_t \Leftarrow \mathbf{x}_t$ (R_t)
Predictor
Measurement Fusion
$\widehat{\mathbf{x}}_{t|t-1} = F_t\widehat{\mathbf{x}}_{t-1|t-1} + B_t\mathbf{u}_t$
$\Sigma_{t|t-1} = F_t\Sigma_{t-1|t-1}F_t^{\mathrm{T}} + Q_t$
$K_t = \Sigma_{t|t-1}(\Sigma_{t|t-1} + R_t)^{-1}$
$\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + K_t(\mathbf{z}_t - \widehat{\mathbf{x}}_{t|t-1})$
$\Sigma_{t|t} = (I - K_t)\Sigma_{t|t-1}$
$\mathbf{z}_t \Leftarrow \mathbf{x}_t$ (R_t)
Predictor
Measurement Fusion
$\widehat{\mathbf{x}}_{t|t-1} = F_t\widehat{\mathbf{x}}_{t-1|t-1} + B_t\mathbf{u}_t$
$\Sigma_{t|t-1} = F_t\Sigma_{t-1|t-1}F_t^{\mathrm{T}} + Q_t$
$K_t = \Sigma_{t|t-1}H_t^{\mathrm{T}}(H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + R_t)^{-1}$
$\widehat{\mathbf{x}}_{t|t} = \widehat{\mathbf{x}}_{t|t-1} + K_t(\mathbf{z}_t - H_t\widehat{\mathbf{x}}_{t|t-1})$
$\Sigma_{t|t} = (I - K_tH_t)\Sigma_{t|t-1}$
$\mathbf{z}_t \Leftarrow H_t\mathbf{x}_t$ (R_t)

(a) Discrete-time dynamical system.

(b) Dynamical system with uncertainty.

(c) Implementation of the dataflow diagram (b).

(d) Implementation of the dataflow diagram (b) for systems with partial observability.

the expectation $E[y|x{=}x_1]$) and return this as the estimate for y if $x{=}x_1$. This requires knowing the joint distribution of x and y, which may not always be available.

In some problems, we can assume that there is an unknown linear relationship between **x** and **y** and that uncertainty comes from noise. Therefore, we can use a technique similar to the ordinary least squares (OLS) method to estimate this linear relationship, and use it to return the best estimate of y for any given value of x. In Figure 5, we see that although there are many points (x_1, y), the y values are clustered around the line as shown in the figure so the value $\hat{y}_1$ is a reasonable estimate for the value of y corresponding to x_1. This line, called the *best linear unbiased estimator* (BLUE), is the analog of ordinary least squares (OLS) for distributions.

Computing BLUE. Consider the estimator $\hat{\mathbf{y}}_{A,\mathrm{b}}(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$. We choose A and **b** so that this is an unbiased estimator with minimal *MSE*. The "^" over the **y** is notation that indicates that we are computing an estimate for **y**.

Theorem 4. *Let*

$$\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \sim p(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}).$$

The estimator $\hat{\boldsymbol{y}}_{A,\boldsymbol{b}}(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$ *for estimating the value of* **y** *for a given value of* **x** *is an unbiased estimator with minimal MSE if*

$$\boldsymbol{b} = \boldsymbol{\mu}_{\mathrm{y}} - A(\boldsymbol{\mu}_{\mathrm{x}})$$

$$A = \Sigma_{yx}\Sigma_{xx}^{-1}$$

The proof of Theorem 4 is straight forward. For an unbiased estimator, $E[\hat{\mathbf{y}}] = E[\mathbf{y}]$. This implies that $\mathbf{b}{=}\boldsymbol{\mu}_{\mathrm{y}}{-}A(\boldsymbol{\mu}_{\mathrm{x}})$ so an unbiased estimator is of the form $\hat{\mathbf{y}}_A(\mathbf{x}) = \boldsymbol{\mu}_{\mathrm{y}} + A(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{x}})$. Note this is equivalent to asserting the BLUE line must pass through the point $(\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\mu}_{\mathrm{y}})$. Setting the derivative of $MSE_A(\hat{\mathbf{y}}_A)$ with respect to A to zero[22] and solving for A, we find that the best linear unbiased estimator is

$$\hat{\mathbf{y}} = \boldsymbol{\mu}_{\mathrm{y}} + \Sigma_{\mathrm{yx}}\Sigma_{\mathrm{xx}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\mathrm{x}}) \qquad (31)$$

This equation can be understood intuitively as follows. If we have no information about **x** and **y**, the best we can do is the estimate $(\boldsymbol{\mu}_{\mathrm{x}}, \boldsymbol{\mu}_{\mathrm{y}})$, which lies on the BLUE line. However, if we know that **x** has a particular value $\mathbf{x}_1$, we can use the correlation between **y** and **x** to estimate a better value for **y** from the difference $(\mathbf{x}_1{-}\boldsymbol{\mu}_{\mathrm{x}})$. Note that if $\sum_{\mathrm{yx}} = 0$ (that is, **x** and **y** are uncorrelated), the best estimate of **y** is just $\boldsymbol{\mu}_{\mathrm{y}}$, so knowing the value of **x** does not give us any additional information about **y** as one would expect. In Figure 5, this corresponds to the case when the BLUE line is parallel to the x-axis. At the other extreme, suppose that **y** and **x** are functionally related so $\mathbf{y} = C\mathbf{x}$. In that case, it is easy to see that $\sum_{\mathrm{yx}} = C\sum_{\mathrm{xx}}$, so $\hat{\mathbf{y}} = C\mathbf{x}$ as expected. In Figure 5, this corresponds to the case when the confidence ellipse shrinks down to the BLUE line.

Equation 31 is a generalization of ordinary least squares in the sense that if we compute the relevant means and variances of a set of discrete data (x_i, y_i) and substitute into Equation 31, we get the same line that is obtained by using OLS.

Kalman Filters for Linear Systems

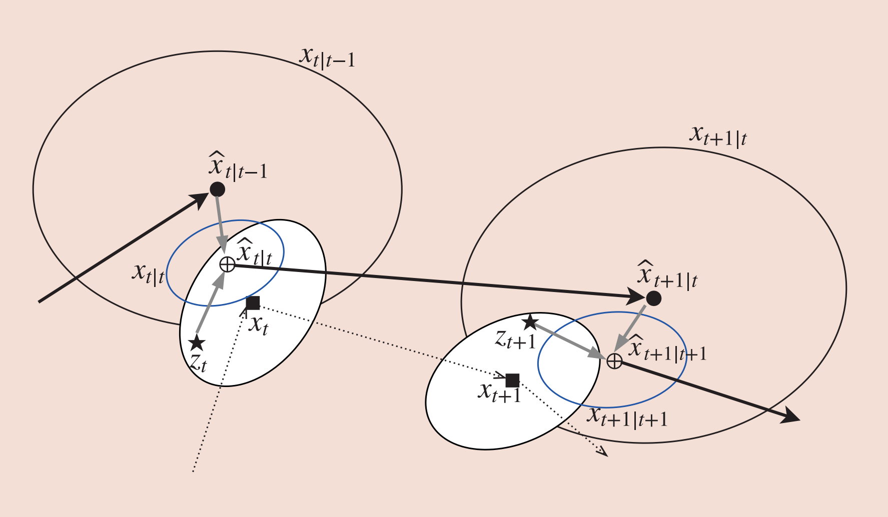
We now apply the algorithms for optimal fusion of vector estimates (Figure 4) and the BLUE estimator (Theorem 4) to the particular problem of state estimation in linear systems, which is the classical application of Kalman filtering.

Figure 6a shows how the evolution of the state of such a system over time can be computed if the initial state $\mathbf{x}_0$ and the model of the system dynamics are known precisely. Time advances in discrete steps. The state of the system at any time step is a function of the state of the system at the previous time step and the control inputs applied to the system during that interval. This is usually expressed by an equation of the form $\mathbf{x}_t = f_t(\mathbf{x}_{t-1}, \mathbf{u}_t)$ where $\mathbf{u}_t$ is the control input. The function f_t is nonlinear in the general case, and can be different for different steps. If the system is linear, the relation for state evolution over time can be written as $\mathbf{x}_t = F_t\mathbf{x}_{t-1} + B_t\mathbf{u}_t$, where F_t and B_t are time-dependent matrices that can be determined from the physics of the system. Therefore, if the initial state $\mathbf{x}_0$ is known exactly and the system dynamics are modeled perfectly by the F_t and B_t matrices, the evolution of the state over time can be computed precisely as shown in Figure 6a.

In general, however, we may not know the initial state exactly, and the system dynamics and control inputs may not be known precisely. These inaccuracies may cause the state computed by the model to diverge unacceptably from the actual state over time. To avoid this, we can make measurements of the state after each time step. If these measurements were exact, there would of course be no need to model the system dynamics. However, in general, the measurements themselves are imprecise.

Kalman filtering was invented to solve the problem of state estimation in such systems. Figure 6b shows the dataflow of the computation, and we use it to introduce standard terminology. An estimate of the initial state, denoted by $\hat{\mathbf{x}}_{0|0}$, is assumed to be available. At each time step t=1, 2, .., the system model is used to provide an estimate of the state at time t using information from time t–1. This step is called *prediction* and the estimate that it provides is called the

Figure 7. Illustration of Kalman filtering.



a priori estimate and denoted by $\hat{\mathbf{x}}_{t|t-1}$. The *a priori* estimate is then fused with $\mathbf{z}_t$, the state estimate obtained from the measurement at time *t*, and the result is the a posteriori state estimate at time *t*, denoted by $\hat{\mathbf{x}}_{t|t}$. This *a posteriori* estimate is used by the model to produce the *a priori* estimate for the next time step and so on. As described here, the *a priori* and *a posteriori* estimates are the means of certain random variables; the covariance matrices of these random variables are shown within parentheses each estimate in Figure 6b, and these are used to weight estimates when fusing them.

We first present the state evolution model and *a priori* state estimation. Then we discuss how state estimates are fused if an estimate of the entire state can be obtained by measurement. Finally, we discuss how to address this problem when only a portion of the state can be measured directly.

State evolution model and prediction. The evolution of the state over time is described by a series of random variables $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$

- The random variable $\mathbf{x}_0$ captures the likelihood of different initial states.
- The random variables at successive time steps are related by the following linear model:

$$\mathbf{x}_t = F_t\mathbf{x}_{t-1} + B_t\mathbf{u}_t + \mathbf{w}_t \qquad (32)$$

Here, $\mathbf{u}_t$ is the control input, which is assumed to be deterministic, and $\mathbf{w}_t$ is a zero-mean noise term that models all the uncertainty in the system. The covariance matrix of $\mathbf{w}_t$ is denoted by Q_t, and the noise terms in different time steps are assumed to be uncorrelated to each other (such as, $E[\mathbf{w}_i\mathbf{w}_j]=0$ if $i\neq j$) and to $\mathbf{x}_0$.

For estimation, we have a random variable $\mathbf{x}_{0|0}$ that captures our belief about the likelihood of different states at time t=0, and two random variables $\mathbf{x}_{t|t-1}$ and $\mathbf{x}_{t|t}$ at each time step t = 1, 2, ... that capture our beliefs about the likelihood of different states at time *t* before and after fusion with the measurement, respectively. The mean and covariance matrix of a random variable $\mathbf{x}_{i|j}$ are denoted by $\hat{\mathbf{x}}_{i|j}$ and $\Sigma_{i|j}$, respectively. We assume $E[\hat{\mathbf{x}}_{0|0}] = E[\mathbf{x}_0]$ (no bias).

Prediction essentially uses $\mathbf{x}_{t-1|t-1}$ as a proxy for $\mathbf{x}_{t-1}$ in Equation 32 to determine $\mathbf{x}_{t|t-1}$ as shown in Equation 33.

Figure 8. Computation of a posteriori estimate.

(i) The *a priori* estimate of the observable part of the state is $H_t\hat{\mathbf{x}}_{t|t-1}$ and its covariance is $H_t\Sigma_{t|t-1}H_t^{\mathrm{T}}$. Using Equation 21 to fuse it with the measurement gives the *a posteriori* estimate of the observable part of the state: $H_t\hat{\mathbf{x}}_{t|t} = H_t\hat{\mathbf{x}}_{t|t-1} + H_t\Sigma_{t|t-1}H_t^{\mathrm{T}}(H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + R_t)^{-1}(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1})$. Let

$$K_t{=}\Sigma_{t|t-1}H_t^{\mathrm{T}}(H_t\Sigma_{t|t-1}H_t^{\mathrm{T}} + R_t)^{-1}.$$

The *a posteriori* estimate of the observable state can be written in terms of K_t as follows:

$$H_t\hat{\mathbf{x}}_{t|t} = H_t\hat{\mathbf{x}}_{t|t-1} + H_tK_t(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1}) \qquad (36)$$

(ii) The *a priori* estimate of the hidden state is $C_t\hat{\mathbf{x}}_{t|t-1}$. The covariance between the hidden portion $C_t\mathbf{x}_{t|t-1}$ and the observable portion $H_t\mathbf{x}_{t|t-1}$ is $C_t\Sigma_{t|t-1}H_t^{\mathrm{T}}$. The difference between the *a priori* estimate and *a posteriori* estimate of $H_t\mathbf{x}$ is $H_tK_t(\mathbf{z}_t{-}H\hat{\mathbf{x}}_{t|t-1})$. Therefore the *a posteriori* estimate of the hidden portion of the state is obtained directly from Equation 31:

$$\begin{aligned} C_t\hat{\mathbf{x}}_{t|t} &= C_t\hat{\mathbf{x}}_{t|t-1} + (C_t\Sigma_{t|t-1}H_t^{\mathrm{T}})(H_t\Sigma_{t|t-1}H_t^{\mathrm{T}})^{-1}H_tK_t(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1}) \\ &= C_t\hat{\mathbf{x}}_{t|t-1} + C_tK_t(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1}) \end{aligned} \qquad (37)$$

(iii) Putting the *a posteriori* estimates (36) and (37) together,

$$\begin{pmatrix} H_t \\ C_t \end{pmatrix}\hat{\mathbf{x}}_{t|t} = \begin{pmatrix} H_t \\ C_t \end{pmatrix}\hat{\mathbf{x}}_{t|t-1} + \begin{pmatrix} H_t \\ C_t \end{pmatrix}K_t(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1}) \qquad (38)$$

Since $\begin{pmatrix} H_t \\ C_t \end{pmatrix}$ is invertible, it can be canceled from the left and right hand sides, giving the equation

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + K_t(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1}) \qquad (39)$$

(iv) To compute $\Sigma_{t|t}$, Equation 39 can be rewritten as $\hat{\mathbf{x}}_{t|t} = (I - K_tH_t)\hat{\mathbf{x}}_{t|t-1} + K_t\mathbf{z}_t$. Since $\mathbf{x}_{t|t-1}$ and $\mathbf{z}_t$ are uncorrelated, it follows from Lemma 2 that

$$\Sigma_{t|t} = (I - K_tH_t)\Sigma_{t|t-1}(I - K_tH_t)^{\mathrm{T}} + K_tR_tK_t^{\mathrm{T}} \qquad (40)$$

Substituting the value of K_t and simplifying, we get

$$\Sigma_{t|t} = (I - K_tH_t)\Sigma_{t|t-1} \qquad (41)$$

$$\mathbf{x}_{t|t-1} = F_t\mathbf{x}_{t-1|t-1} + B_t\mathbf{u}_t + \mathbf{w}_t \tag{33}$$

For state estimation, we need only the mean and covariance matrix of $\mathbf{x}_{t|t-1}$. The predictor box in Figure 6 computes these values; the covariance matrix is obtained from Lemma 2 under the assumption that $\mathbf{w}_t$ is uncorrelated with $\mathbf{x}_{t-1|t-1}$, which is justified here.

Fusing complete observations of the state. If the entire state can be measured at each time step, the imprecise measurement at time t is modeled as follows:

$$\mathbf{z}_t = \mathbf{x}_t + \mathbf{v}_t \tag{34}$$

where $\mathbf{v}_t$ is a zero-mean noise term with covariance matrix R_t. The noise terms in different time steps are assumed to be uncorrelated with each other (such as, $E[\mathbf{v}_i\mathbf{v}_j]$ is zero if $i \neq j$) as well as with $\mathbf{x}_{0|0}$ and all $\mathbf{w}_k$. A subtle point here is that $\mathbf{x}_t$ in this equation is the actual state of the system at time t (that is, a particular realization of the random variable $\mathbf{x}_t$), so variability in $\mathbf{z}_t$ comes only from $\mathbf{v}_t$ and its covariance matrix R_t.

Therefore, we have two imprecise estimates for the state at each time step $t = 1, 2, \ldots$, the *a priori* estimate from the predictor $(\hat{\mathbf{x}}_{t|t-1})$ and the one from the measurement $(\mathbf{z}_t)$. If $\mathbf{z}_t$ is uncorrelated to $\mathbf{x}_{t|t-1}$, we can use Equations 20–22 to fuse the estimates as shown in Figure 6c.

The assumptions that (i) $\mathbf{x}_{t-1|t-1}$ is uncorrelated with $\mathbf{w}_t$, which is used in prediction, and (ii) $\mathbf{x}_{t|t-1}$ is uncorrelated with $\mathbf{z}_t$, which is used in fusion, are easily proved to be correct by induction on t, using Lemma 2(ii). Figure 6b gives the intuition: $\mathbf{x}_{t|t-1}$ for example is an affine function of the random variables $\mathbf{x}_{0|0}, \mathbf{w}_1, \mathbf{v}_1, \mathbf{w}_2, \mathbf{v}_2, \ldots, \mathbf{w}_t$, and is therefore uncorrelated with $\mathbf{v}_t$ (by assumption about $\mathbf{v}_t$ and Lemma 2(ii)) and hence with $\mathbf{z}_t$.

Figure 7 shows the computation pictorially using confidence ellipses to illustrate uncertainty. The dotted arrows at the bottom of the figure show the evolution of the state, and the solid arrows show the computation of the *a priori* estimates and their fusion with measurements.

Fusing partial observations of the state. In some problems, only a portion of the state can be measured directly. The observable portion of the state is specified formally using a full row-rank matrix H_t called the *observation matrix*: if the state is $\mathbf{x}$, what is observable is $H_t\mathbf{x}$. For example, if the state vector has two components and only the first component is observable, H_t can be $[1\ 0]$. In general, the H_t matrix can specify a linear relationship between the state and the observation, and it can be time-dependent. The imprecise measurement model introduced in Equation 34 becomes:

$$\mathbf{z}_t = H_t\mathbf{x}_t + \mathbf{v}_t \tag{35}$$

The hidden portion of the state can be specified using a matrix C_t, which is an orthogonal complement of H_t. For example, if $H_t = [1\ 0]$, one choice for C_t is $[0\ 1]$.

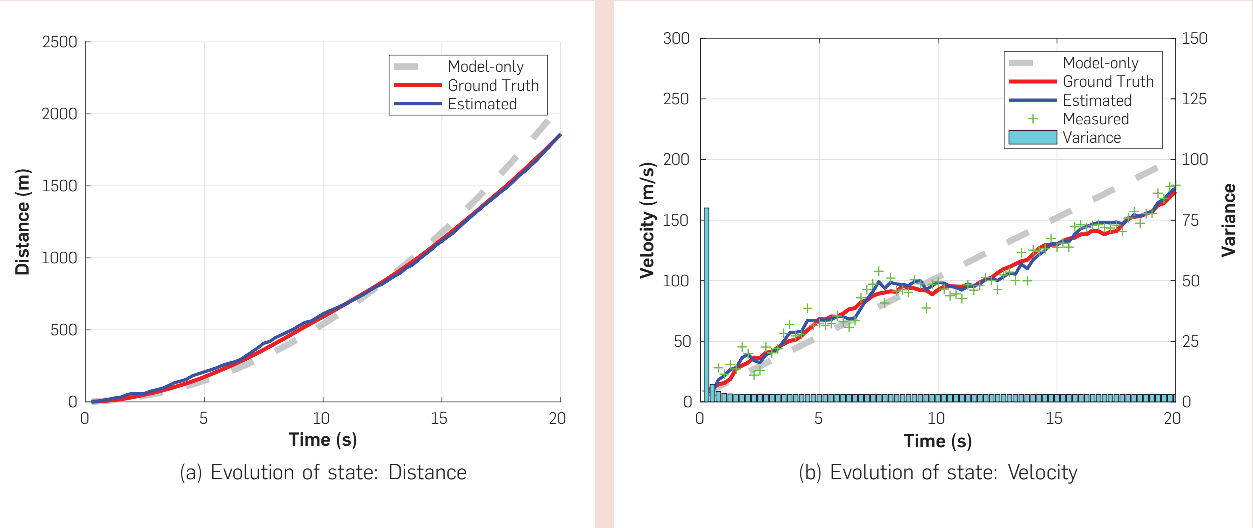
Figure 6d shows the computation for this case. The fusion phase can be understood intuitively in terms of the following steps.

i. The observable part of the *a priori* estimate of the state $(H_t\hat{\mathbf{x}}_{t|t-1})$ is fused with the measurement $(\mathbf{z}_t)$, using Equations 20–22. The quantity $(\mathbf{z}_t - H_t\hat{\mathbf{x}}_{t|t-1})$ is called the *innovation*. The result is the *a posteriori* estimate of the observable state $(H_t\hat{\mathbf{x}}_{t|t})$.
ii. The BLUE of Theorem 4 is used to obtain the *a posteriori* estimate of the hidden state $(C_t\hat{\mathbf{x}}_{t|t})$ by adding to the *a priori* estimate of the hidden state $(C_t\hat{\mathbf{x}}_{t|t-1})$ a value obtained from the product of the covariance between $H_t\mathbf{x}_{t|t-1}$ and $C_t\mathbf{x}_{t|t-1}$ and the difference between $H_t\hat{\mathbf{x}}_{t|t-1}$ and $H_t\hat{\mathbf{x}}_{t|t}$.
iii. The *a posteriori* estimates of the observable and hidden portions of the state are composed to produce the *a posteriori* estimate of the entire state $(\hat{\mathbf{x}}_{t|t-1})$.

The actual implementation produces the final result directly without going through these steps as shown in Figure 6d, but these incremental steps are useful for understanding how all this works, and their implementation is shown in more detail in Figure 8.

Figure 6d puts all this together. In the literature, this dataflow is referred to as Kalman filtering. Unlike in Equations 18 and 21, the

Figure 9. Estimates of the object's state over time.



(a) Evolution of state: Distance

(b) Evolution of state: Velocity

Kalman gain is not a dimensionless value here. If $H_t = I$, the computations in Figure 6d reduce to those of Figure 6c as expected.

Equation 39 shows that the *a posteriori* state estimate is a linear combination of the *a priori* state estimate ($\hat{\mathbf{x}}_{t|t-1}$) and the measurement ($\mathbf{z}_t$). The optimality of this linear unbiased estimator is shown in the Appendix. It was shown earlier that incremental fusion of scalar estimates is optimal. The dataflow of Figures 6(c,d) computes the *a posteriori* state estimate at time t by incrementally fusing measurements from the previous time steps, and this incremental fusion can be shown to be optimal using a similar argument.

Example: falling body. To demonstrate the effectiveness of the Kalman filter, we consider an example in which an object falls from the origin at time t=0 with an initial speed of 0 m/s and an expected constant acceleration of 9.8 m/s^2 due to gravity. Note that acceleration in reality may not be constant due to factors such as wind, and air friction.

The state vector of the object contains two components, one for the distance from the origin $s(t)$ and one for the velocity $v(t)$. We assume that only the velocity state can be measured at each time step. If time is discretized in steps of 0.25 seconds, the difference equation for the dynamics of the system is easily shown to be the following:

$$\begin{pmatrix} v_n \\ s_n \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.25 & 1 \end{pmatrix} \begin{pmatrix} v_{n-1} \\ s_{n-1} \end{pmatrix} + \begin{pmatrix} 0 & 0.25 \\ 0 & 0.5 \times 0.25^2 \end{pmatrix} \begin{pmatrix} 0 \\ 9.8 \end{pmatrix}$$

where we assume $\begin{pmatrix} v_0 \\ s_0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $\Sigma_0 = \begin{pmatrix} 80 & 0 \\ 0 & 10 \end{pmatrix}$.

The gray lines in Figure 9 show the evolution of velocity and distance with time according to this model. Because of uncertainty in modeling the system dynamics, the actual evolution of the velocity and position will be different in practice. The red lines in Figure 9 show one trajectory for this evolution, corresponding to a Gaussian noise term with covariance $\begin{pmatrix} 2 & 2.5 \\ 2.5 & 4 \end{pmatrix}$ in Equation 32 (because this noise term is random, there are many trajectories for the evolution, and we are just showing one of them).

We have shown that Kalman filtering for state estimation in linear systems can be derived from two elementary ideas: optimal linear estimators for fusing uncorrelated estimates and best linear unbiased estimators for correlated variables.

The red lines correspond to "ground truth" in our example.

The green points in Figure 9b show the noisy measurements of velocity at different time steps, assuming the noise is modeled by a Gaussian with variance 8. The blue lines show the *a posteriori* estimates of the velocity and position. It can be seen that the *a posteriori* estimates track the ground truth quite well even when the ideal system model (the gray lines) is inaccurate and the measurements are noisy. The cyan bars in the right figure show the variance of the velocity at different time steps. Although the initial variance is quite large, application of Kalman filtering is able to reduce it rapidly in few time steps.

Discussion. We have shown that Kalman filtering for state estimation in linear systems can be derived from two elementary ideas: optimal linear estimators for fusing uncorrelated estimates and best linear unbiased estimators for correlated variables. This is a different approach to the subject than the standard presentations in the literature. One standard approach is to use Bayesian inference. The other approach is to assume that the *a posteriori* state estimator is a linear combination of the form $A_t\hat{\mathbf{x}}_{t|t-1} + B_t\mathbf{z}_t$, and then find the values of A_t and B_t that produce an unbiased estimator with minimum *MSE*. We believe that the advantage of the presentation given here is that it exposes the concepts and assumptions that underlie Kalman filtering.

Most presentations in the literature also begin by assuming that the noise terms $\mathbf{w}_t$ in the state evolution equation and $\mathbf{v}_t$ in the measurement are Gaussian. Although some presentations[1,10] use properties of Gaussians to derive the results in Figure 3, these results do not depend on distributions being Gaussians. Gaussians however enter the picture in a deeper way if one considers *nonlinear* estimators. It can be shown that if the noise terms are not Gaussian, there may be nonlinear estimators whose *MSE* is lower than that of the linear estimator presented in Figure 6d. However, if the noise is Gaussian, this linear estimator is as good as any unbiased nonlinear estimator (that is, the linear estimator is a *minimum variance unbiased estimator*

(MVUE)). This result is proved using the Cramer-Rao lower bound.[24]

Extension to Nonlinear Systems

The *extended Kalman filter* (EKF) and *unscented Kalman filter* (UKF) are heuristic approaches to using Kalman filtering for nonlinear systems. The state evolution and measurement equations for nonlinear systems with additive noise can be written as follows; in these equations, f and h are nonlinear functions.

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_t) + \mathbf{w}_t \qquad (42)$$

$$\mathbf{z}_t = h(\mathbf{x}_t) + \mathbf{v}_t \qquad (43)$$

Intuitively, the EKF constructs linear approximations to the nonlinear functions f and h and applies the Kalman filter equations, whereas the UKF constructs approximations to probability distributions and propagates these through the nonlinear functions to construct approximations to the posterior distributions.

EKF. Examining Figure 6d, we see that the *a priori* state estimate in the predictor can be computed using the system model: $\hat{\mathbf{x}}_{t|t-1} = f(\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{u}_t)$. However, as the system dynamics and measurement equations are nonlinear, it is not clear how to compute the co-variance matrices for the *a priori* estimate and the measurement. In the EKF, these matrices are computed by linearizing Equations 42 and 43 using the Taylor series expansions for the nonlinear functions f and h. This requires computing the following *Jacobians*,[e] which play the role of F_t and H_t in Figure 6d.

$$F_t = \left.\frac{\partial f}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{t-1|t-1}, \mathbf{u}_t} \qquad H_t = \left.\frac{\partial h}{\partial \mathbf{x}}\right|_{\hat{\mathbf{x}}_{t|t-1}}$$

The EKF performs well in some applications such as navigation systems and GPS.[28]

UKF. When the system dynamics and observation models are highly nonlinear, the unscented Kalman filter (UKF)[15] can be an improvement over the EKF. The UKF is based on the *unscented transformation*, which is a method for computing the statistics of a random variable **x** that undergoes a nonlinear transformation ($\mathbf{y} = g(\mathbf{x})$). The random variable **x** is sampled using a carefully chosen set of *sigma points* and these sample points are propagated through the nonlinear function g. The statistics of **y** are estimated using a weighted sample mean and covariance of the posterior sigma points. The UKF tends to be more robust and accurate than the EKF but has higher computation overhead due to the sampling process.

Conclusion

In this article, we have shown that two concepts—optimal linear estimators for fusing uncorrelated estimates and best linear unbiased estimators for correlated variables—provide the underpinnings for Kalman filtering. By combining these ideas, standard results on Kalman filtering for linear systems can be derived in an intuitive and straightforward way that is simpler than other presentations of this material in the literature. This approach makes clear the assumptions that underlie the optimality results associated with Kalman filtering and should make it easier to apply Kalman filtering to problems in computer systems. **C**

Acknowledgments

This research was supported by NSF grants 1337281, 1406355, and 1618425, and by DARPA contracts FA8750-16-2-0004 and FA8650-15-C-7563. We are indebted to K. Mani Chandy (Caltech), Ivo Babuska (UT Austin,) and Augusto Ferrante (Padova, Italy) for their valuable feedback.

[e] The Jacobian matrix is the matrix of all first order partial derivatives of a vector-valued function.

References

1. Babb, T. How a Kalman filter works, in pictures | bzarg. 2018. https://www.bzarg.com/p/how-a-kalman-filter-works-in-pictures/. Accessed: 2018-11-30
2. Balakrishnan, A.V. *Kalman Filtering Theory*. Optimization Software, Inc., Los Angeles, CA, USA, 1987.
3. Barker, A.L., Brown, D.E., Martin, W.N. *Bayesian Estimation and the Kalman Filter*. Technical Report. Charlottesville, VA, USA, 1994.
4. Becker, A. Kalman filter overview. https://www.kalmanfilter.net/default.aspx. 2018. Accessed: 2018-11-08.
5. Bergman, K. Nanophotonic interconnection networks in multicore embedded computing. In *2009 IEEE/LEOS Winter Topicals Meeting Series* (2009), 6–7.
6. Cao, L., Schwartz, H.M. Analysis of the Kalman filter based estimation algorithm: An orthogonal decomposition approach. *Automatica 1*, 40 (2004), 5–19.
7. Chui, C.K., Chen, G. *Kalman Filtering: With Real-Time Applications*, 5th edn. Springer Publishing Company, Incorporated, 2017.
8. Eubank, R.L. *A Kalman Filter Primer (Statistics: Textbooks and Monographs)*. Chapman & Hall/CRC, 2005.
9. Evensen, G. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
10. Faragher, R. Understanding the basis of the Kalman filter via a simple and intuitive derivation. *IEEE Signal Process. Mag. 5*, 29 (2012), 128–132.
11. Grewal, M.S., Andrews, A.P. *Kalman Filtering: Theory and Practice with MATLAB*, 4th edn. Wiley-IEEE Press, 2014.
12. Hess, A.-K., Rantzer, A. Distributed Kalman filter algorithms for self-localization of mobile devices. In *Proceedings of the 13th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '10, 2010, 191–200.
13. Imes, C., Kim, D.H.K., Maggio, M., Hoffmann, H. POET: A portable approach to minimizing energy under soft real-time constraints. In *21st IEEE Real-Time and Embedded Technology and Applications Symposium*, April 2015, 75–86.
14. Imes, C., Hoffmann, H. Bard: A unified framework for managing soft timing and power constraints. In *International Conference on Embedded Computer Systems: Architectures, Modeling and Simulation (SAMOS)*, 2016.
15. Julier, S.J., Uhlmann, J.K. Unscented filtering and nonlinear estimation. *Proc. IEEE 3*, 92 (2004), 401–422.
16. Kitanidis, P.K. Unbiased minimum-variance linear state estimation. *Automatica 6*, 23 (1987), 775–778.
17. Lindquist, A., Picci, G. *Linear Stochastic Systems*. Springer-Verlag, 2017.
18. Maybeck, P.S. *Stochastic Models, Estimation, and Control*, volume 3. Academic Press, 1982.
19. Mendel, J.M. *Lessons in Estimation Theory for Signal Processing, Communications, and Control*. Pearson Education, 1995.
20. Nagarajan, K., Gans, N., Jafari, R. Modeling human gait using a Kalman filter to measure walking distance. In *Proceedings of the 2nd Conference on Wireless Health*, WH '11 (New York, NY, USA, 2011). ACM, 34:1–34:2.
21. Nakamura, E.F., Loureiro, A.A.F., Frery, A.C. Information fusion for wireless sensor networks: methods, models, and classifications. *ACM Comput. Surv. 3*, 39 (2007).
22. Petersen, K.B., Pedersen, M.S. The Matrix Cookbook. http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274. November 2012. Version 20121115.
23. Pothukuchi, R.P., Ansari, A., Voulgaris, P., Torrellas, J. Using multiple input, multiple output formal control to maximize resource efficiency in architectures. In *Proceedings of the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)* (2016), IEEE, 658–670.
24. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, 37 (1945), 81–89.
25. Rhudy, M.B., Salguero, R.A., Holapa, K. A Kalman filtering tutorial for undergraduate students. *Int. J. Comp. Sci. Eng. Surv.* (1), 8 (2017).
26. Sengupta, S.K. Fundamentals of statistical signal processing: Estimation theory. *Technometrics* (4), 37 (1995), 465–466.
27. Souza, É.L., Nakamura, E.F., Pazzi, R.W. Target tracking for sensor networks: A survey. *ACM Comput. Surv.* (2), 49 (2016), 30:1–30:31.
28. Thrun, S., Burgard, W., Fox, D. *Probabilistic Robotics (Intelligent Robotics and Autonomous Agents)*. The MIT Press, 2005.
29. Welch, G., Bishop, G. *An Introduction to the Kalman Filter*. Technical Report. Chapel Hill, NC, USA, 1995.
30. Pei, Y., Biswas, S., Fussell, D.S., Pingali, K. An Elementary Introduction to Kalman Filtering. *ArXiv e-prints*. October 2017.

The online appendix for this article can be found at http://dl.acm.org/citation.cfm?doid=3363294&picked=formats

Yan Pei (ypei@cs.utexas.edu) is a graduate research assistant in the Department of Computer Science at the University of Texas, Austin, TX, USA.

Swarnendu Biswas (swarnendu@cse.iitk.ac.in) is an assistant professor in the Department of Computer Science and Engineering at the Indian Institute of Technology, Kanpur, India.

Donald S. Fussell (fussell@cs.utexas.edu) is the Trammell Crow Regents Professor in the Department of Computer Science at the University of Texas, Austin, TX, USA.

Keshav Pingali (pingali@cs.utexas.edu) is a professor in the Department of Computer Science at the University of Texas, Austin, and the W.A."Tex" Moncrief Chair of Computing in the UT Oden Institute of Computational Engineering and Science, Austin, TX, USA.

research highlights

DOI:10.1145/3361564

To view the accompanying paper, visit doi.acm.org/10.1145/3361566 rh

Technical Perspective
A Whitebox Solution for Blackbox-Like Behaviors

By David G. Andersen

DEEP NEURAL NETWORKS (DNNs) are rapidly becoming an indispensable part of the computing toolbox, with particular success in helping to bridge the messy analog world into forms we can process with more conventional computing techniques (image and speech recognition, as some of the most obvious examples).

The price we pay, however, is inscrutability: DNNs behave like black boxes, without clearly explainable logic for their functioning. Admitting for the moment that most complex software systems are also approximately impossible to fully reason about, we have—and continue to develop—methods for formally reasoning about and extensively testing critical components. Almost nothing equivalent exists for DNNs. This is particularly worrying precisely because of the power of DNNs to allow us to extend computing into domains previously inaccessible. In at least one area of medical diagnostics—identifying diabetic retinopathy—DNN-based approaches already match expert human performance, but we have little experience yet to help us understand what kind of bugs those systems may fall prey to when deployed in the real world.

DeepXplore brings a software testing perspective to DNNs and, in doing so, creates the opportunity for enormous amounts of follow-on work in several ways. Much of the prior work in finding errors in DNNs focused on finding individual adversarial modifications of images, but without the explicit focus on a diversity of computational paths taken by the DNN to achieve them. The metric introduced in DeepXplore—neuron coverage—is an analogue of the code coverage metric traditionally used in software testing. This metric has utility beyond the techniques used in DeepXplore; security bug hunting, for example, has found coverage-guided fuzzing to be a powerful and effective technique, and the neuron coverage metric and its derivatives can enable similar approaches in the DNN context.

I often tell students, when first starting to learn about research, that they should keep an eye out for the papers in an area that everyone else claims to have beaten: Those are the papers that stimulated other researchers. DeepXplore will be such a paper. Its specific metrics and constraints on example generation are unlikely to be the final word in DNN testing, but the work that follows will exist because of researchers seeing these ideas and trying to improve upon them. The core framework from DeepXplore will likely endure: Establish an effective coverage metric based upon the numerical values obtained by the activations of the neural network and use a constrained search procedure to maximize coverage with respect to that metric. C

I often tell students to keep an eye out for the papers in an area that everyone else claims to have beaten: Those are the papers that stimulated other researchers. DeepXplore will be such a paper.

David G. Andersen is a professor in the computer science department at Carnegie Mellon University, Pittsburgh, PA, USA, and is CTO of BrdgAI.

DOI:10.1145/3361566

DeepXplore: Automated Whitebox Testing of Deep Learning Systems

By Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana

Abstract

Deep learning (DL) systems are increasingly deployed in safety- and security-critical domains such as self-driving cars and malware detection, where the correctness and predictability of a system's behavior for corner case inputs are of great importance. Existing DL testing depends heavily on manually labeled data and therefore often fails to expose erroneous behaviors for rare inputs.

We design, implement, and evaluate DeepXplore, the first white-box framework for systematically testing real-world DL systems. First, we introduce neuron coverage for measuring the parts of a DL system exercised by test inputs. Next, we leverage multiple DL systems with similar functionality as cross-referencing oracles to avoid manual checking. Finally, we demonstrate how finding inputs for DL systems that both trigger many differential behaviors and achieve high neuron coverage can be represented as a joint optimization problem and solved efficiently using gradient-based search techniques.

DeepXplore efficiently finds thousands of incorrect corner case behaviors (e.g., self-driving cars crashing into guard rails and malware masquerading as benign software) in state-of-the-art DL models with thousands of neurons trained on five popular datasets such as ImageNet and Udacity self-driving challenge data. For all tested DL models, on average, DeepXplore generated one test input demonstrating incorrect behavior within one second while running only on a commodity laptop. We further show that the test inputs generated by DeepXplore can also be used to retrain the corresponding DL model to improve the model's accuracy by up to 3%.

1. INTRODUCTION

Over the past few years, Deep Learning (DL) has made tremendous progress, achieving or surpassing human-level performance for a diverse set of tasks in many application domains. These advances have led to widespread adoption and deployment of DL in security- and safety-critical systems such as self-driving cars,[1] malware detection,[4] and aircraft collision avoidance systems.[6]

This wide adoption of DL techniques presents new challenges as the predictability and correctness of such systems are of crucial importance. Unfortunately, DL systems, despite their impressive capabilities, often demonstrate unexpected or incorrect behaviors in corner cases for several reasons such as biased training data and overfitting of the models. In safety- and security-critical settings, such incorrect behaviors can lead to disastrous consequences such as a fatal collision of a self-driving car. For example, a Google self-driving car recently crashed into a bus because it expected the bus to yield under a set of rare conditions but the bus did not.[a]

A Tesla car in autopilot crashed into a trailer because the autopilot system failed to recognize the trailer as an obstacle due to its "white color against a brightly lit sky" and the "high ride height".

[b] Such corner cases were not part of Google's or Tesla's test set and thus never showed up during testing.

Therefore, DL systems, just like traditional software, must be tested systematically for different corner cases to detect and fix ideally any potential flaws or undesired behaviors. This presents a new system problem as automated and systematic testing of large-scale, real-world DL systems with thousands of neurons and millions of parameters for all corner cases is extremely challenging.

The standard approach for testing DL systems is to gather and manually label as much real-world test data as possible. Some DL systems such as Google self-driving cars also use simulation to generate synthetic training data. However, such simulation is completely unguided as it does not consider the internals of the target DL system. Therefore, for the large input spaces of real-world DL systems (e.g., all possible road conditions for a self-driving car), none of these approaches can hope to cover more than a tiny fraction (if any at all) of all possible corner cases.

Recent works on adversarial deep learning[3] have demonstrated that carefully crafted synthetic images by adding minimal perturbations to an existing image can fool state-of-the-art DL systems. The key idea is to create synthetic images such that they get classified by DL models differently than the original picture but still look the same to the human eye. Although such adversarial images expose some erroneous behaviors of a DL model, the main restriction of such an approach is that it must limit its perturbations to tiny invisible changes and require ground truth labels. Moreover, just like other forms of existing DL testing, the adversarial images only cover a small part (52.3%) of DL system's logic as shown in Section 5. In essence, the current machine learning testing practices for finding incorrect corner cases are analogous to finding bugs in traditional software by using

[a] http://www.theverge.com/2016/2/29/11134344/google-self-driving-car-crash-report

[b] https://electrek.co/2016/07/01/understanding-fatal-tesla-accident-autopilot-nhtsa-probe/

The original version of this paper was published in *Proceedings of the 26th Symposium on Operating Systems Principles* (Shanghai, China, Oct. 28–31, 2017), 1–18.

test inputs with low code coverage and thus are unlikely to find many erroneous cases.

The key challenges in automated systematic testing of large-scale DL systems are twofold: (1) how to generate inputs that trigger different parts of a DL system's logic and uncover different types of erroneous behaviors, and (2) how to identify erroneous behaviors of a DL system without manual labeling/checking. This paper describes and highlights how we design and build DeepXplore to address both challenges.

First, we introduce the concept of neuron coverage for measuring the parts of a DL system's logic exercised by a set of test inputs based on the number of neurons activated by the inputs. At a high level, neuron coverage of DL systems is similar to code coverage of traditional systems, a standard empirical metric for measuring the amount of code exercised by an input in the traditional software. However, code coverage itself is not a good metric for estimating coverage of DL systems as most rules in DL systems, unlike traditional software, are not written manually by a programmer but rather learned from training data. In fact, we find that for most of the DL systems that we tested, even a single randomly picked test input was able to achieve 100% code coverage, whereas the neuron coverage was less than 10%.

Next, we show how multiple DL systems with similar functionality (e.g., self-driving cars by Google and Tesla, and GM) can be used as cross-referencing oracles to identify erroneous corner cases without providing ground truth labels which require huge manual labeling effort. For example, if one self-driving car decides to turn left whereas others turn right for the same input, one of them is likely to be incorrect. Such techniques have been applied successfully in the past for detecting logic bugs without manual specifications in a wide variety of traditional software.[2] In this paper, we demonstrate how differential testing can be applied to DL systems.

Finally, we demonstrate how the problem of generating test inputs that maximize neuron coverage of a DL system while also exposing as many differential behaviors (i.e., differences between multiple similar DL systems) as possible can be formulated as a joint optimization problem. Unlike traditional programs, Deep Neural Networks (DNNs) used by DL systems are differentiable. Therefore, their gradients with respect to inputs can be calculated accurately given whitebox access to the corresponding model. In this paper, we show how these gradients can be used to efficiently solve the joint optimization problem for large-scale real-world DL systems.

We design, implement, and evaluate DeepXplore, to the best of our knowledge, the first efficient whitebox testing framework for large-scale DL systems. In addition to maximizing neuron coverage and behavioral differences between DL systems, DeepXplore also supports adding custom constraints by the users for simulating different types of realistic inputs (e.g., different types of lighting and occlusion for images/videos). We demonstrate that DeepXplore efficiently finds thousands of unique incorrect corner case behaviors (e.g., self-driving cars crashing into guard rails) in 15 state-of-the-art DL models trained using five real-world datasets such as Udacity self-driving car challenge data, image data from ImageNet and MNIST, Android malware data from Drebin, and PDF malware data from Contagio/VirusTotal. For all of the tested DL models, on average, DeepXplore generated one test input demonstrating incorrect behavior within one second while running on a commodity laptop. The inputs generated by DeepXplore achieved 34.4 and 33.2% higher neuron coverage on average than the same number of randomly picked inputs and adversarial inputs,[3] respectively. We further show that the test inputs generated by DeepXplore can be used to retrain the corresponding DL model to improve classification accuracy as well as identify potentially polluted training data. We achieve up to 3% improvement in classification accuracy by retraining a DL model on inputs generated by DeepXplore compared to retraining on the same number of random or adversarial inputs.

A number of follow-up papers after DeepXplore have expanded the idea of whitebox testing for domain-specific transformations in self-driving cars[13] and developed exhaustive black box testing techniques for a variety of common transformations.[11] Besides, the metric of neuron coverage has also been extended in TensorFlow as an open-source tool by the Google Brain team.[10] Beyond testing, we have also studied and proposed more rigorous verification techniques leveraging interval arithmetic to certify the robustness of neural networks[15, 16] (Figure 1).

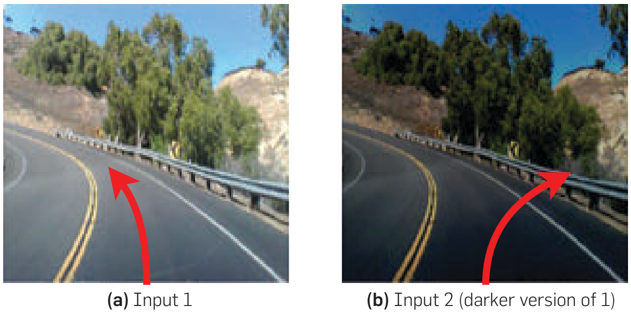
2. BACKGROUND

2.1. DL systems

We define a DL system to be any software system that includes at least one Deep Neural Network (DNN) component. Note that some DL systems might comprise solely DNNs (e.g., self-driving car DNNs predicting steering angles without any manual rules), whereas others may have some DNN components interacting with other traditional software components to produce the final output.

As shown in Figure 2, the development process of the DNN components of a DL system is fundamentally different from traditional software development. Unlike traditional software, where the developers directly specify the logic of the system, the DNN components learn their rules

Figure 1. An example of erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a).



(a) Input 1

(b) Input 2 (darker version of 1)

automatically from data. The developers of DNN components can indirectly influence the rules learned by a DNN by modifying the training data, features, and the model's architectural details (e.g., number of layers).

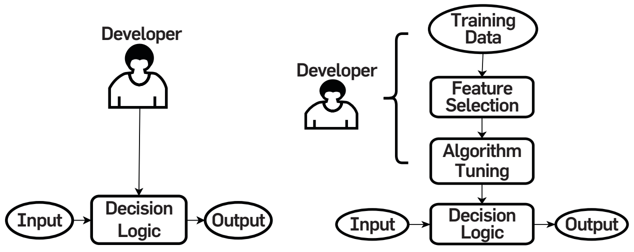
2.2. DNN architecture

DNNs are inspired by human brains with millions of interconnected neurons. They are known for their amazing ability to automatically identify and extract the relevant high-level features from raw inputs without any human guidance besides labeled training data. In recent years, DNNs have surpassed human performance in many application domains due to increasing availability of large datasets, specialized hardware, and efficient training algorithms.

A DNN consists of multiple *layers*, each containing multiple *neurons* as shown in Figure 3. A *neuron* is an individual computing unit inside a DNN that applies an *activation function* on its inputs and passes the result to other connected neurons (see Figure 3). The common activation functions include sigmoid, hyperbolic tangent, or ReLU (Rectified Linear Unit). A DNN usually has at least three (often more) layers: one input, one output, and one or more hidden layers. Each neuron in one layer has directed connections to the neurons in the next layer. The numbers of neurons in each layer and the connections between them vary significantly across DNNs. Overall, a DNN can be defined mathematically as a multi-input, multi-output parametric function *F* composed of many parametric subfunctions representing different neurons.

Each connection between the neurons in a DNN is bound to a *weight* parameter characterizing the strength of the connection between the neurons. For supervised learning, the weights of the connections are learned during training by minimizing a cost function over the training data via gradient descent.

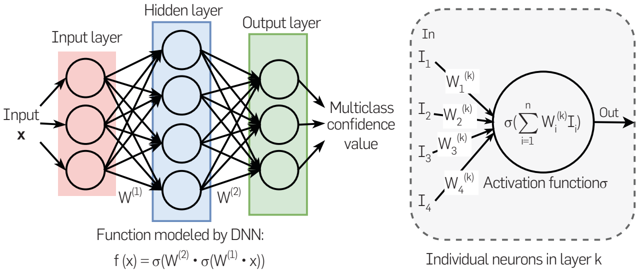
Figure 2. Comparison between traditional and ML system development processes. Developers specify clear logic of the system, whereas DNN learns the logic from training data.



Traditional software development

ML system development

Figure 3. A simple DNN and the computations performed by each neuron.



Individual neurons in layer k

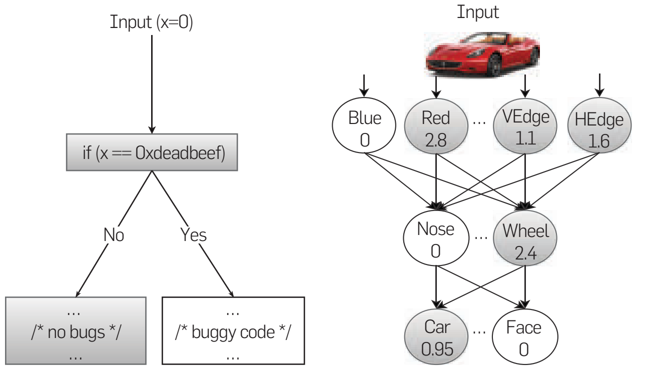
Each layer of the network transforms the information contained in its input to a higher-level representation of the data. For example, consider a pretrained network as shown in Figure 4b for classifying images into two categories: human faces and cars. The first few hidden layers transform the raw pixel values into low-level texture features such as edges or colors and feed them to the deeper layers.[18] The last few layers, in turn, extract and assemble the meaningful high-level abstractions such as noses, eyes, wheels, and headlights to make the classification decision.

2.3. Limitations of existing DNN testing

Expensive labeling effort. Existing DNN testing techniques require prohibitively expensive human effort to provide correct labels/actions for a target task (e.g., self-driving a car, image classification, and malware detection). For complex and high-dimensional real-world inputs, human beings, even domain experts, often have difficulty in efficiently performing a task correctly for a large dataset. For example, consider a DNN designed to identify potentially malicious executable files. Even a security professional will have trouble determining whether an executable is malicious or benign without executing it. However, executing and monitoring a malware inside a sandbox incur significant performance overhead and therefore make manual labeling significantly harder to scale to a large number of inputs.

Low test coverage. None of the existing DNN testing schemes even try to cover different rules of the DNN. Therefore, the test inputs often fail to uncover different erroneous behaviors of a DNN. For example, DNNs are often tested by simply dividing a whole dataset into two random parts—one for training and the other for testing. The testing set in such cases may only exercise a small subset of all rules learned by a DNN. Recent results involving adversarial evasion attacks against DNNs have demonstrated the existence of some corner cases where DNN-based image

Figure 4. Comparison between program flows of a traditional program and a neural network. The nodes in gray denote the corresponding basic blocks or neurons that participated while processing an input.



(a) A program with a rare branch

(b) A DNN for detecting cars and faces classifiers (with state-of-the-art performance on randomly picked testing sets) still incorrectly classify synthetic images generated by adding humanly imperceptible perturbations to a test image.[3] However, the adversarial inputs, similar to random test inputs, also only cover a small part of the rules learned by a DNN as they are not designed to maximize coverage. Moreover, they are also inherently limited to small imperceptible perturbations around a test input as larger perturbations will visually change the input and therefore will require manual inspection to ensure correctness of the DNN's decision.

Problems with low-coverage DNN tests. To better understand the problem of low test coverage of rules learned by a DNN, we provide an analogy to a similar problem in testing traditional software. Figure 4 shows a side-by-side comparison of how a traditional program and a DNN handle inputs and produce outputs. Specifically, the figure shows the *similarity between traditional software and DNNs*: in software program, each statement performs a certain operation to transform the output of previous statement(s) to the input to the following statement(s), whereas in DNN, each neuron transforms the output of previous neuron(s) to the input of the following neuron(s). Of course, unlike traditional software, DNNs do not have explicit branches but a neuron's influence on the downstream neurons decreases as the neuron's output value gets lower. A lower output value indicates less influence and vice versa. When the output value of a neuron becomes zero, the neuron does not have any influence on the downstream neurons.

As demonstrated in Figure 4a, the problem of low coverage in testing traditional software is obvious. In this case, the buggy behavior will never be seen unless the test input is `0xdeadbeef`. The chances of randomly picking such a value are very small. Similarly, low-coverage test inputs will also leave different behaviors of DNNs unexplored. For example, consider a simplified neural network, as shown in Figure 4b, that takes an image as an input and classifies it into two different classes: cars and faces. The text in each neuron (represented as a node) denotes the object or property that the neuron detects,[c] and the number in each neuron is the real value outputted by that neuron. The number indicates how confident the neuron is about its output. Note that randomly picked inputs are highly unlikely to set high output values for the unlikely combination of neurons. Therefore, many incorrect DNN behaviors will remain unexplored even after performing a large number of random tests. For example, if an image causes neurons labeled as "Nose" and "Red" to produce high output values and the DNN misclassifies the input image as a car, such a behavior will never be seen during regular testing as the chances of an image containing a red nose (e.g., a picture of a clown) are very small.

3. OVERVIEW

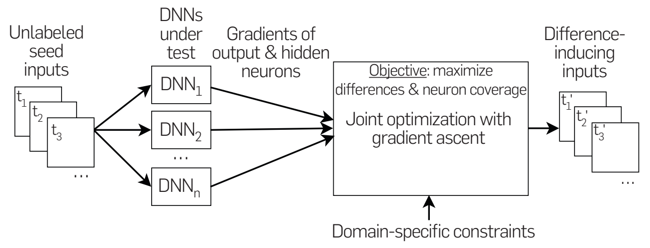
In this section, we provide a general overview of DeepXplore, our whitebox framework for systematically testing DNNs for erroneous corner case behaviors. The main components of DeepXplore are shown in Figure 5. DeepXplore takes unlabeled test inputs as seeds and generates new tests that cover a large number of neurons (i.e., activates them to a value above a customizable threshold) while causing the tested DNNs to behave differently. Specifically, DeepXplore solves a joint optimization problem that maximizes both differential behaviors and neuron coverage. Note that both goals are crucial for thorough testing of DNNs and finding diverse erroneous corner case behaviors. High neuron coverage alone may not induce many erroneous behaviors, whereas just maximizing different behaviors might simply identify different manifestations of the same underlying root cause.

DeepXplore also supports enforcing of custom domain-specific constraints as part of the joint optimization process. For example, the value of an image pixel has to be between 0 and 255. Such domain-specific constraints can be specified by the users of DeepXplore to ensure that the generated test inputs are valid and realistic.

We designed an algorithm for efficiently solving the joint optimization problem mentioned above using gradient ascent. First, we compute the gradient of the *outputs* of the neurons in both the output and hidden layers with the *input value* as a variable and the *weight parameter* as a constant. Such gradients can be computed efficiently for most DNNs. Note that DeepXplore is designed to operate on pretrained DNNs. The gradient computation is efficient because our whitebox approach has access to the pretrained DNNs' weights and the intermediate neuron values. Next, we iteratively perform gradient ascent to modify the test input toward maximizing the objective function of the joint optimization problem described above. Essentially, we perform a gradient-guided local search starting from the seed inputs and find new inputs that maximize the desired goals. Note that, at a high level, our gradient computation is similar to the backpropagation performed during the training of a DNN, but the key difference is that, unlike our algorithm, backpropagation treats the *input value* as a constant and the *weight parameter* as a variable.

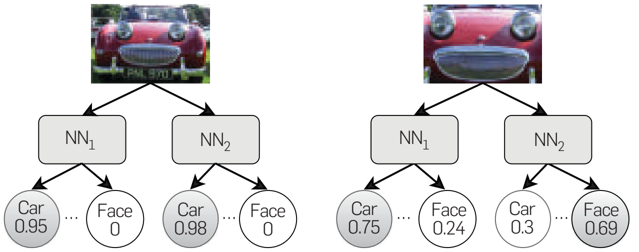
A working example. We use Figure 6 as an example to show how DeepXplore generates test inputs. Consider that we have two DNNs to test—both perform similar tasks, that is, classifying images into cars or faces, as shown in Figure 6, but they are trained independently with different datasets and parameters. Therefore, the DNNs will learn similar but slightly different classification rules. Let us also assume that

Figure 5. DeepXplore workflow.


Unlabeled seed inputs
t_1
t_2
t_3
...
DNNs under test
DNN$_1$
DNN$_2$
...
DNN$_n$
Gradients of output & hidden neurons
Objective: maximize differences & neuron coverage
Joint optimization with gradient ascent
Domain-specific constraints
Difference-inducing inputs
t_1'
t_2'
t_3'
...

[c] Note that one cannot always map each neuron to a particular task, i.e., detecting specific objects/properties. Figure 4b simply highlights that different neurons often tend to detect different features.

Figure 6. Inputs inducing different behaviors in two similar DNNs.



(a) DNNs produce same output **(b)** DNNs produce different output

we have a seed test input, the image of a red car, which both DNNs identify as a car as shown in Figure 6a.

DeepXplore tries to maximize the chances of finding differential behavior by modifying the input, that is, the image of the red car, towards maximizing its probability of being classified as a car by one DNN but minimizing corresponding probability of the other DNN. DeepXplore also tries to cover as many neurons as possible by activating (i.e., causing a neuron's output to have a value greater than a threshold) inactive neurons in the hidden layer. We further add domain-specific constraints (e.g., ensure the pixel values are integers within 0 and 255 for image input) to make sure that the modified inputs still represent real-world images. The joint optimization algorithm will iteratively perform a gradient ascent to find a modified input that satisfies all of the goals described above. DeepXplore will eventually generate a set of test inputs where the DNNs' outputs differ, for example, one DNN thinks it is a car, whereas the other thinks it is a face as shown in Figure 6b.

4. METHODOLOGY

In this section, we provide a brief technical description of our algorithm. The details can be found in the original paper. First, we define and explain the concepts of neuron coverage and gradient for DNNs. Next, we describe how the testing problem can be formulated as a joint optimization problem. Finally, we provide the gradient-based algorithm for solving the joint optimization problem.

4.1. Definitions

Neuron coverage. We define neuron coverage of a set of test inputs as the ratio of the number of unique activated neurons for all test inputs and the total number of neurons in the DNN.[d] We consider a neuron activated if its output is greater than a threshold (e.g., 0).

More formally, let us assume that all neurons of a DNN are represented by the set $N = \{n_1, n_2, ...\}$, all test inputs are represented by the set $T = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ...\}$, and $out(n, \boldsymbol{x})$ is a function that returns the output value of neuron n in the DNN for a given test input $\boldsymbol{x}$. Note that the bold $\boldsymbol{x}$ signifies that $\boldsymbol{x}$ is a vector. Let t represent the threshold for considering a neuron to be activated. In this setting, neuron coverage can be defined as follows.

$$NCov(T,\boldsymbol{x}) = \frac{|\{n \mid \forall \boldsymbol{x} \in T, out(n,\boldsymbol{x}) > t\}|}{|N|}$$

To demonstrate how neuron coverage is calculated in practice, consider the DNN as shown in Figure 4b. The neuron coverage (with threshold 0) for the input picture of the red car as shown in Figure 4b will be $5/8 = 0.625$.

Gradient. The gradients or forward derivatives of the outputs of neurons of a DNN with respect to the input are well known in deep learning literature. They have been extensively used both for crafting adversarial examples and visualizing/understanding DNNs.[18] We provide a brief definition here for completeness and refer interested readers to[18] for more details.

Let $\boldsymbol{\theta}$ and $\boldsymbol{x}$ represent the parameters and the test input of a DNN, respectively. The parametric function performed by a neuron can be represented as $y = f(\boldsymbol{\theta}, \boldsymbol{x})$ where f is a function that takes $\boldsymbol{\theta}$ and $\boldsymbol{x}$ as input and output y. Note that y can be the output of any neuron defined in the DNN (e.g., neuron from output layer or intermediate layers). The gradient of $f(\boldsymbol{\theta}, \boldsymbol{x})$ with respect to input $\boldsymbol{x}$ can be defined as:

$$\boldsymbol{G} = \nabla_x f(\boldsymbol{\theta}, \boldsymbol{x}) = \partial y / \partial \boldsymbol{x} \qquad (1)$$

The computation inside f is essentially a sequence of stacked functions that compute the input from previous layers and forward the output to next layers. Thus, $\boldsymbol{G}$ can be calculated by utilizing the chain rule in calculus, that is, by computing the layer-wise derivatives starting from the layer of the neuron that outputs y until reaching the input layer that takes $\boldsymbol{x}$ as the input. Note that the dimension of the gradient $\boldsymbol{G}$ is identical to that of the input $\boldsymbol{x}$.

4.2. DeepXplore algorithm

The main advantage of the test input generation process for a DNN over traditional software is that the test generation process, once defined as an optimization problem, can be solved efficiently using gradient ascent. In this section, we describe the details of the formulation and find solutions to the optimization problem. Note that solutions to the optimization problem can be efficiently found for DNNs as the gradients of the objective functions of DNNs, unlike traditional software, can be easily computed.

As discussed earlier in Section 3, the objective of the test generation process is to maximize both the number of observed differential behaviors and the neuron coverage while preserving domain-specific constraints provided by the users. Below, we define the objectives of our joint optimization problem formally and explain the details of the algorithm for solving it.

Maximizing differential behaviors. The first objective of the optimization problem is to generate test inputs that can induce different behaviors in the tested DNNs, that is, different DNNs will classify the same input into different classes. Suppose we have n DNNs $F_{k\in 1..n}: \boldsymbol{x} \rightarrow \boldsymbol{y}$, where F_k is the function modeled by the kth neural network. $\boldsymbol{x}$ represents the input and $\boldsymbol{y}$ represents the output class probability vectors.

[d] Neuron coverage can be defined in many different ways other than that defined in this paper. We refer readers to other follow-up papers for details on different definitions.

Given an arbitrary $\boldsymbol{x}$ as seed that gets classified to the same class by all DNNs, our goal is to modify $\boldsymbol{x}$ such that the modified input $\boldsymbol{x'}$ will be classified differently by at least one of the n DNNs.

Let $F_k(\boldsymbol{x})[c]$ be the class probability that F_k predicts $\boldsymbol{x}$ to be c. We randomly select one neural network F_j and maximize the following objective function:

$$obj_1(\boldsymbol{x}) = \sum\nolimits_{k \neq j} F_k(\boldsymbol{x})[c] - \lambda_1 \cdot F_j(\boldsymbol{x})[c] \qquad (2)$$

where λ_1 is a parameter to balance the objective terms between the DNNs' $F_{k \neq j}$ that maintain the same class outputs as before and the DNN F_j that produce different class outputs. As all of $F_{k \in 1..n}$ are differentiable, Equation 2 can be maximized using gradient ascent by iteratively changing $\boldsymbol{x}$ based on the computed gradient: $\frac{\partial obj_1(\boldsymbol{x})}{\partial \boldsymbol{x}}$.

Maximizing neuron coverage. The second objective is to generate inputs that maximize neuron coverage. We achieve this goal by iteratively picking inactivated neurons and modifying the input such that the output of that neuron goes above the neuron activation threshold. Let us assume that we want to maximize the output of a neuron n, that is, we want to maximize $obj_2(\boldsymbol{x}) = f_n(\boldsymbol{x})$ such that $f_n(\boldsymbol{x}) > t$, where t is the neuron activation threshold, and we write $f_n(\boldsymbol{x})$ as the function modeled by neuron n that takes $\boldsymbol{x}$ (the original input to the DNN) as the input and produce the output of neuron n (as defined in Equation 1). We can again leverage the gradient ascent mechanism as $f_n(\boldsymbol{x})$ is a differentiable function whose gradient is $\frac{\partial f_n(\boldsymbol{x})}{\partial \boldsymbol{x}}$.

Note that we can also potentially jointly maximize multiple neurons simultaneously, but we choose to activate one neuron at a time in this algorithm for clarity.

Joint optimization. We jointly maximize obj_1 and f_n described above and maximize the following function:

$$obj_{joint} = \left(\sum\nolimits_{i \neq j} F_i(\boldsymbol{x})[c] - \lambda_1 F_j(\boldsymbol{x})[c]\right) + \lambda_2 \cdot f_n(\boldsymbol{x}) \qquad (3)$$

where λ_2 is a parameter for balancing between the two objectives and n is the inactivated neuron that we randomly pick at each iteration. As all terms of obj_{joint} are differentiable, we jointly maximize them using gradient ascent by modifying $\boldsymbol{x}$.

Domain-specific constraints. One important aspect of the optimization process is that the generated test inputs need to satisfy several domain-specific constraints to be physically realistic. In particular, we want to ensure that the changes applied to $\boldsymbol{x_i}$ during the ith iteration of gradient ascent process satisfy all the domain-specific constraints for all i. For example, for a generated test image $\boldsymbol{x}$, the pixel values must be within a certain range (e.g., 0–255).

Although some such constraints can be efficiently embedded into the joint optimization process using the Lagrange Multipliers similar to those used in support vector machines, we found that the majority of them cannot be easily handled by the optimization algorithm. Therefore, we designed a simple rule-based method to ensure that the generated tests satisfy the custom domain-specific constraints. As the seed input $\boldsymbol{x_{seed}} = \boldsymbol{x_0}$ always satisfies the constraints by definition, our technique must ensure that after the ith ($i > 0$) iteration of gradient ascent, $\boldsymbol{x_i}$ still satisfies the constraints. Our algorithm ensures this property by modifying the gradient $\boldsymbol{grad}$ such that $\boldsymbol{x_{i+1}} = \boldsymbol{x_i} + s \cdot \boldsymbol{grad}$ still satisfies the constraints (s is the step size in the gradient ascent).

For discrete features, we round the gradient to an integer. For DNNs handling visual input (e.g., images), we add different spatial restrictions such that only part of the input images is modified. A detailed description of the domain-specific constraints that we implemented can be found in Section 5.2.

Hyperparameters. To summarize, there are four major hyperparameters that control different aspects of DeepXplore as described below. (1) λ_1 balances the objectives between minimizing one DNN's prediction for a certain label and maximizing the rest of DNNs' predictions for the same label. Larger λ_1 puts higher priority on lowering the prediction value/confidence of a particular DNN, whereas smaller λ_1 puts more weight on maintaining the other DNNs' predictions. (2) λ_2 provides balance between finding differential behaviors and neuron coverage. Larger λ_2 focuses more on covering different neurons, whereas smaller λ_2 generates more difference-inducing test inputs. (3) s controls the step size used during iterative gradient ascent. Larger s may lead to oscillation around the local optimum, whereas smaller s may need more iterations to reach the objective. (4) t is the threshold to determine whether each individual neuron is activated or not. Finding inputs that activate a neuron becomes increasingly harder as t increases.

5. EXPERIMENTAL SETUP

5.1. Test datasets and DNNs

We adopt 5 popular public datasets with different types of data—MNIST, ImageNet, Driving, Contagio/VirusTotal, and Drebin—and then evaluate DeepXplore on 3 DNNs for each dataset (i.e., a total of 15 DNNs). We provide a summary of the five datasets and the corresponding DNNs in Table 1. The detailed description can be found in the full paper. All the evaluated DNNs are either pretrained (i.e., we use public weights reported by previous researchers) or trained by us using public real-world architectures to achieve comparable performance to that of the state-of-the-art models for the corresponding dataset. For each dataset, we used DeepXplore to test three DNNs with different architectures.

5.2. Domain-specific constraints

As discussed earlier, to be useful in practice, we need to ensure that the generated tests are valid and realistic by applying domain-specific constraints. For example, generated images should be physically producible by a camera. Similarly, generated PDFs need to follow the PDF specification to ensure that a PDF viewer can open the test file. Below we describe two major types of domain-specific constraints (i.e., image and file constraints) that we use in this paper. **Image constraints (MNIST, ImageNet, and Driving).** DeepXplore used three different types of constraints for simulating different environmental conditions of images: (1) lighting effects for simulating different intensities of lights, (2) occlusion by a single small rectangle for simulating an attacker potentially blocking some parts of a camera,

Table 1. Details of the DNNs and datasets used to evaluate DeepXplore.

Dataset	Dataset description	DNN description	DNN name	# of neurons	Architecture	Reported Acc.	Our Acc.
MNIST	Hand-written digits	LeNet variations	MNI_C1	52	LeNet-1, LeCun et al. [8]	98.3%	98.33%
			MNI_C2	148	LeNet-4, LeCun et al. [8]	98.9%	98.59%
			MNI_C3	268	LeNet-5, LeCun et al. [8]	99.05%	98.96%
Imagenet	General images	State-of-the-art image classifiers from ILSVRC	IMG_C1	14,888	VGG-16, Simonyan et al. [12]	92.6%**	92.6%**
			IMG_C2	16,168	VGG-19, Simonyan et al. [12]	92.7%**	92.7%**
			IMG_C3	94,059	ResNet50, He et al. [5]	96.43%**	96.43%**
Driving	Driving video frames	Nvidia DAVE self-driving systems	DRV_C1	1,560	Dave-orig [1]	N/A	99.91%$^{\#}$
			DRV_C2	1,560	Dave-norminit$^{\#\#}$	N/A	99.94%$^{\#}$
			DRV_C3	844	Dave-dropout^{++}	N/A	99.96%$^{\#}$
Contagio/ Virustotal	PDFs	PDF malware detectors	PDF_C1	402	<200, 200>$^{+}$	98.5%$^{-}$	96.15%
			PDF_C2	602	<200, 200, 200>$^{+}$	98.5%$^{-}$	96.25%
			PDF_C3	802	<200, 200, 200, 200>$^{+}$	98.5%$^{-}$	96.47%
Drebin	Android apps	Android app malware detectors	APP_C1	402	<200, 200>$^{+}$, Grosse et al. [4]	98.92%	98.6%
			APP_C2	102	<50, 50>$^{+}$, Grosse et al. [4]	96.79%	96.82%
			APP_C3	212	<200, 10>$^{+}$, Grosse et al. [4]	92.97%	92.66%

** Top-5 test accuracy; we exactly match the reported performance as we use the pretrained networks.
# We report 1-MSE (mean squared error) as the accuracy because steering angle is a continuous value.
+ <x,y,...> denotes three hidden layers with *x* neurons in first layer, *y* neurons in second layer, etc.
– Accuracy using SVM as reported by Šrndic et al. [14].
## https://github.com/jacobgil/keras-steering-angle-visualizations.
++ https://github.com/navoshta/behavioral-cloning.

and (3) occlusion by multiple tiny black rectangles for simulating effects of dirt on camera lens.

Other constraints (Drebin and Contagio/VirusTotal). For Drebin dataset, DeepXplore enforces a constraint that only allows modifying features related to the Android manifest file and thus ensures that the application code is unaffected. Moreover, DeepXplore only allows adding features (changing from zero to one) but does not allow deleting features (changing from one to zero) from the manifest files to ensure that no application functionality is changed due to insufficient permissions. Thus, after computing the gradient, DeepXplore only modifies the manifest features whose corresponding gradients are greater than zero. For Contagio/VirusTotal dataset, we follow the restrictions on each feature as described by Šrndic and Laskkov.[14]

6. RESULTS

6.1. Summary

Table 2 summarizes the numbers of erroneous behaviors found by DeepXplore for each tested DNN while using 2000 randomly selected seed inputs from the corresponding test sets. Note that as the testing set has a similar number of samples for each class, these randomly-chosen 2000 samples also follow that distribution. The hyperparameters for these experiments, as shown in Table 2, are empirically chosen to maximize both the rate of finding difference-inducing inputs as well as the neuron coverage.

For the experimental results shown in Figure 7, we apply three domain-specific constraints (lighting effects, occlusion by a single rectangle, and occlusion by multiple rectangles) as described in Section 5.2. For all other experiments involving vision-related tasks, we only use the lighting effects as the domain-specific constraints. For all malware-related experiments, we apply all the relevant domain-specific constraints described in Section 5.2. We use the hyperparameters listed in Table 2 in all the experiments unless otherwise specified. Figure 7 shows some difference-inducing inputs generated by DeepXplore for MNIST, ImageNet, and Driving dataset along with the corresponding erroneous behaviors. Table 3 (Drebin) and Table 4 (Contagio/VirusTotal) show two sample difference-inducing inputs generated by DeepXplore that caused erroneous behaviors in the tested DNNs. We highlight the differences between the seed input features and the features modified by DeepXplore. Note that we only list the top three modified features due to space limitations.

Table 2. Number of difference-inducing inputs found by DeepXplore for each tested DNN obtained by randomly selecting 2000 seeds from the corresponding test set for each run.

DNN name	Hyperparams				# Differences found
	λ_1	λ_2	*s*	*t*	
MNI_C1	1	0.1	10	0	1073
MNI_C2					1968
MNI_C3					827
IMG_C1	1	0.1	10	0	1969
IMG_C2					1976
IMG_C3					1996
DRV_C1	1	0.1	10	0	1720
DRV_C2					1866
DRV_C3					1930
PDF_C1	2	0.1	0.1	0	1103
PDF_C2					789
PDF_C3					1253
APP_C1	1	0.5	N/A	0	2000
APP_C2					2000
APP_C3					2000

6.2. Benefits of neuron coverage

In this subsection, we evaluate how effective neuron coverage is in measuring the comprehensiveness of DNN testing.

Figure 7. The first row shows the seed test inputs and the second row shows the difference-inducing test inputs generated by DeepXplore. The left three columns show results under different lighting effects, the middle three are using single occlusion box, and the right three are using black rectangles as the transformation constraints. For each type of transformation (three pairs of images), the images from left to right are from self-driving car, MNIST, and ImageNet.

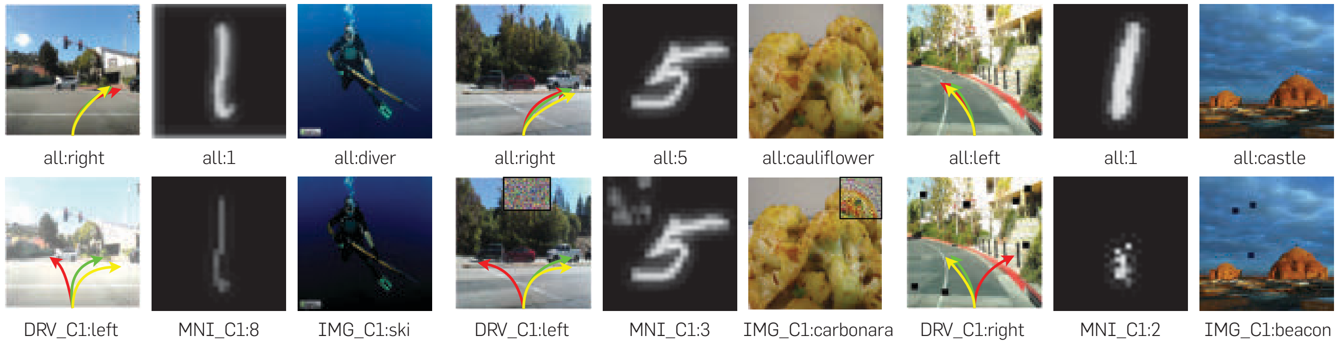


Table 3. The features added to the manifest file for generating two malware inputs that Android app classifiers (Drebin) incorrectly mark as benign.

input 1	**feature**	*feature:: bluetooth*	*activity:: .SmartAlertTerms*	**service_ receiver:: .rrltpsi**
	before	0	0	0
	after	1	1	1
input 2	**feature**	*provider:: xclockprovider*	*permission:: CALL_PHONE*	*provider:: contentprovider*
	before	0	0	0
	after	1	1	1

Table 4. The top-3 most in(de)cremented features for generating two sample malware inputs that PDF classifiers incorrectly mark as benign.

input 1	**feature**	*size*	*count_action*	*count_endobj*
	before	1	0	1
	after	34	21	20
input 2	**feature**	*size*	*count_font*	*author_num*
	before	1	0	10
	after	27	15	5

Table 5. Comparison of code coverage and neuron coverage for 10 randomly selected inputs from the original test set of each DNN.

	Code coverage			**Neuron coverage**		
Dataset	***C1***	***C2***	***C3***	***C1***	***C2***	***C3***
MNIST	100%	100%	100%	32.7%	33.1%	25.7%
ImageNet	100%	100%	100%	1.5%	1.1%	0.3%
Driving	100%	100%	100%	2.5%	3.1%	3.9%
VirusTotal	100%	100%	100%	19.8%	17.3%	17.3%
Drebin	100%	100%	100%	16.8%	10%	28.6%

It has recently been shown that each neuron in a DNN tends to independently extract a specific feature of the input instead of collaborating with other neurons for feature extraction.[18] This finding intuitively explains why neuron coverage is a good metric for DNN testing comprehensiveness. To empirically confirm this observation, we perform two different experiments as described below.

First, we show that neuron coverage is a significantly better metric than code coverage for measuring comprehensiveness of the DNN test inputs. More specifically, we find that a small number of test inputs can achieve 100% code coverage for all DNNs where neuron coverage is actually less than 34%. Second, we evaluate neuron activations for test inputs from different classes. Our results show that inputs from different classes tend to activate more unique neurons than inputs from the same class. Both findings confirm that neuron coverage provides a good estimation of the numbers and types of DNN rules exercised by an input.

Neuron coverage vs. code coverage. We compare both code and neuron coverages achieved by the same number of inputs by evaluating the test DNNs on ten randomly picked testing samples as described in Section 5.1. We measure a DNN's code coverage in terms of the line coverage of the Python code used in the training and testing process. We set the threshold t in neuron coverage 0.75, that is, a neuron is considered covered only if its output is greater than 0.75 for at least one input.

The results, as shown in Table 5, clearly demonstrate that neuron coverage is a significantly better metric than code coverage for measuring DNN testing comprehensiveness. Even 10 randomly picked inputs result in 100% code coverage for all DNNs, whereas the neuron coverage never goes above 34% for any of the DNNs. Moreover, neuron coverage changes significantly based on the tested DNNs and the test inputs. For example, the neuron coverage for the complete MNIST testing set (i.e., 10,000 testing samples) only reaches 57.7, 76.4, and 83.6% for C1, C2, and C3, respectively. In contrast, the neuron coverage for the complete Contagio/Virustotal test set reaches 100%.

Activation of neurons for different classes of inputs. We measure the number of active neurons that are common across the LeNet-5 DNN running on pairs of MNIST inputs of the same and different classes, respectively. In particular, we randomly select 200 input pairs where 100 pairs have the same label (e.g., labeled as 8) and 100 pairs have different labels (e.g., labeled as 8 and 4). Then, we calculate the number of common (overlapped) active neurons for these input pairs. Table 6 confirms our hypothesis that inputs coming from the same class share more activated neurons than those coming from different classes. As inputs from different classes tend to get detected through matching of different DNN rules, our result also confirms that neuron coverage can effectively estimate the numbers of different rules activated during DNN testing.

7. LIMITATIONS AND FUTURE WORKS

Although our results are very encouraging, several other obstacles must be solved to make ML systems more reliable.

First, DeepXplore only considers a small subset of transformations to test the corresponding properties. Although they are arguably more realistic than adversarial perturbations, they still do not fully capture all real-world input distortions. Tian et al. have recently developed a testing tool for autonomous vehicles[13] that considers a wider range of transformations and uses neuron coverage to guide the search for errors. However, testing complex realistic transformations such as simulating shadows from other objects still remains an open problem.

Next, it is challenging to *efficiently* search for error-inducing test cases for arbitrary transformations. DeepXplore efficiently finds error-inducing inputs leveraging the input gradients. However, there are many realistic transformations for which such input gradient information cannot be computed accurately. For example, it is difficult to compute gradients directly to emulate different weather conditions (e.g., snow or rain) for testing self-driving vehicles. There is an emerging area of research that leverages the generative adversarial networks (GANs) to learn differentiable representations of such complex transformations to enable gradient-based search for error-inducing inputs.[9]

Finally, a key limitation of our gradient-based local search is that it does not provide any guarantee about the absence of errors. There has been recent progress on two complementary directions that can provide stronger guarantees than DeepXplore. First, Pei et al. considered a specific subset of transformations where the output space is polynomial in the input image size.[11] Therefore, it is feasible for these transformations to exhaustively enumerate the transformed inputs to verify the absence of errors. Second, several recent works have explored new formal verification techniques for NNs[7,15–17], that can either ensure the absence of adversarial inputs or provide a concrete counterexample for a given network and a test input. However, scaling these techniques to larger networks remains a major challenge.

Table 6. Average number of overlaps among activated neurons for a pair of inputs of the same class and different classes. Inputs of different classes tend to activate different neurons.

	Total neurons	*Avg. no. of activated neurons*	*Avg. overlap*
Diff. class	268	83.6	45.9
Same class	268	84.1	74.2

8. CONCLUSION

We designed and implemented DeepXplore, the first whitebox system for systematically testing DL systems. We introduced a new metric, neuron coverage, for measuring how many rules in a DNN are exercised by a set of inputs. DeepXplore performs gradient ascent to solve a joint optimization that maximizes both neuron coverage and the number of potentially erroneous behaviors. DeepXplore was able to find thousands of erroneous behaviors in 15 state-of-the-art DNNs trained on five real-world datasets. We hope DeepXplore's results and its limitations can encourage and motivate other researchers to work on this challenging but critical and exciting area.

References

1. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316* (2016).
2. Brubaker, C., Jana, S., Ray, B., Khurshid, S., Shmatikov V. Using frankencerts for automated adversarial testing of certificate validation in SSL/TLS implementations. In *Proceedings of the 35th IEEE Symposium on Security and Privacy* (2014).
3. Goodfellow, I., Shlens, J., Szegedy, C. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations* (2015).
4. Grosse, K., Papernot, N., Manoharan, P., Backes, M., McDaniel, P. Adversarial examples for malware detection. In *European Symposium on Research in Computer Security* (2017).
5. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (2016).
6. Julian, K.D., Lopez, J., Brush, J.S., Owen, M.P., Kochenderfer, M.J. Policy compression for aircraft collision avoidance systems. In *Proceedings of the 35th IEEE/AIAA Digital Avionics Systems Conference* (2016).
7. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J. Reluplex: An efficient smt solver for verifying deep neural networks. In *Proceedings of the 29th International Conference on Computer Aided Verification* (2017).
8. LeCun, Y., Cortes, C., Burges, C.J. MNIST handwritten digit database. 2010.
9. Liu, M.-Y., Breuel, T., Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems* (2017).
10. Odena, A., Goodfellow, I. Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. *arXiv preprint arXiv:1807.10875* (2018).
11. Pei, K., Cao, Y., Yang, J., Jana, S. Towards practical verification of machine learning: The case of computer vision systems. *arXiv preprint arXiv:1712.01785* (2017).
12. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations* (2015).
13. Tian, Y., Pei, K., Jana, S., Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering*, ACM (2018), 303–314
14. Šrndic, N., Laskov, P. Practical evasion of a learning-based classifier: a case study. In *Proceedings of the 35th IEEE Symposium on Security and Privacy* (2014).
15. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems* (2018).
16. Wang, S., Pei, K., Whitehouse, J., Yang, J., Jana, S. Formal security analysis of neural networks using symbolic intervals. In *27th USENIX Security Symposium* (2018).
17. Wong, E., Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning* (2018).
18. Yosinski, J., Clune, J., Fuchs, T., Lipson, H. Understanding neural networks through deep visualization. In *2015 ICML Workshop on Deep Learning* (2015).

Kexin Pei, Junfeng Yang, and Suman Jana ([kpei,junfeng,suman]@cs.columbia.edu), Columbia University, USA.

Yinzhi Cao (yinzhi.cao@jhu.edu), Johns Hopkins University, USA.

Watch the authors discuss their work in this exclusive *Communications* video. https://cacm.acm.org/videos/deepxplore

Boston College

Non Tenure-Track Positions in Computer Science

The Computer Science Department of Boston College seeks to fill one or more non-tenure-track teaching positions, as well as shorter-term visiting teaching positions. All applicants should be committed to excellence in undergraduate education and be able to teach a broad variety of undergraduate computer science courses. Faculty in longer-term positions will also participate in the development of new courses that reflect the evolving landscape of the discipline.

Minimum requirements for the title of Assistant Professor of the Practice, and for the title of Visiting Assistant Professor, include a Ph.D. in Computer Science or closely related discipline.

Candidates without a Ph.D. would be eligible for the title of Lecturer or Visiting Lecturer.

We will begin reviewing applications on October 15, 2019 and will continue considering applications until the positions are filled. Applicants should submit a cover letter, CV, and a separate teaching statement and arrange for three confidential letters of recommendation that comment on their teaching performance to be uploaded directly to Interfolio. To apply go to https://apply.interfolio.com/68339. Boston College conducts background checks as part of the hiring process. Information about the university and our department is available at https://www.bc.edu and http:// cs.bc.edu.

Boston College is a Jesuit, Catholic university that strives to integrate research excellence with a foundational commitment to formative liberal arts education. We encourage applications from candidates who are committed to fostering a diverse and inclusive academic community. Boston College is an affirmative action/equal opportunity employer.

Boston College

Tenure-Track Assistant Professor of Computer Science

The Computer Science Department of Boston College seeks a tenure-track Assistant Professor beginning in the 2020-2021 academic year. Successful candidates for the position will be expected to develop strong research programs that can attract external funding in an environment that also values high-quality undergraduate teaching. Outstanding candidates in all areas of Computer Science will be considered, with a preference for those who demonstrate a potential to contribute to cross-disciplinary teaching and research in conjunction with the planned Schiller Institute for Integrated Science and Society at Boston College.

A Ph.D. in Computer Science or a closely related discipline is required. See http://cs.bc.edu and https://www.bc.edu/bc-web/schools/mcas/sites/schiller-institute.html for more information. Application review is ongoing. Boston College conducts background checks as part of the hiring process.

Submit a cover letter, a detailed CV and teaching and research statements. Arrange for three confidential letters of recommendation to be uploaded directly to Interfolio. To apply go to https://apply.interfolio.com/68273.

Boston College is a Jesuit, Catholic university that strives to integrate research excellence with a foundational commitment to formative liberal arts education. We encourage applications from candidates who are committed to fostering a diverse and inclusive academic community. Boston College is an affirmative action/equal opportunity employer.

Columbia University

Junior Faculty Position in the Department of Electrical Engineering

Columbia Engineering is pleased to invite applications for a faculty position in the Department of **Electrical Engineering** at Columbia University in the City of New York. Applications for Junior rank will be considered.

The Electrical Engineering department welcomes applications in all areas of electrical engineering (http://www.ee.columbia.edu/ee-research). Candidates must have a Ph.D. or its professional equivalent by the starting date of the appointment. Applicants for this position must demonstrate the potential to do pioneering research and to teach effectively. The Department is especially interested in qualified candidates who can contribute, through their research, teaching, and/or service, to the diversity and excellence of the academic community.

The successful candidate is expected to contribute to the advancement of their field and the department by developing an original and leading externally funded research program, and to contribute to the undergraduate and graduate educational mission of the Department. Columbia fosters multidisciplinary research and encourages collaborations with academic departments and units across Columbia University. The Department actively participates in the school-wide Engineering for Humanity initiatives that relate to engineering and medicine, autonomous systems, quantum computing and technology, and sustainability.

For additional information and to apply, please see: http://engineering.columbia.edu/faculty-job-opportunities. Applications should be submitted electronically and include the following: curriculum vitae including a publication list, a description of research accomplishments, statements of research and teaching interests and plans, contact information for three experts who can provide letters of recommendation, and up to three pre/reprints of scholarly work. All applications received by December 1st, 2019 will receive full consideration. We will accept and review applications after this date.

Applicants can consult http://www.ee.columbia.edu for more information about the department and http://pa334.peopleadmin.com/postings/4208 for more details on the position and application.

Columbia University is an Equal Opportunity Employer / Disability / Veteran.

Georgia Institute of Technology

Multiple Tenure Track Faculty Positions

The School of Computer Science at the Georgia Institute of Technology (Georgia Tech) invites applications for several tenure track faculty positions at all ranks. We seek candidates in all areas who complement and enhance our current research strengths and are especially interested this year in candidates whose research focus is in the broad area of Theoretical Computer Science.

Georgia Tech is an equal education/employment opportunity institution dedicated to building a diverse community. We strongly encourage applications from women, underrepresented groups, individuals with disabilities, and veterans. Georgia Tech has policies to promote a healthy work-life balance and is aware that attracting faculty may require meeting the needs of two careers.

The School of Computer Science, one of three schools in the top-ten ranked College of Computing, focuses on research that makes computing and communication smart, fast, reliable, and secure, with research groups in computer architecture, databases, machine learning, networking, programming languages, security, software engineering, systems, and theory. Faculty in the school are leaders in a variety of Georgia Tech initiatives, including: the Algorithms and Randomness Center (ARC), the Center for Research into Novel Computing Hierarchies (CRNCH), the Institute for Data Engineering and Science (IDEaS), and the Institute for Information Security and Privacy (IISP). The school is in a period of rapid growth with eight tenure-track Assistant Professors hired in the last two years.

Georgia Tech is a top-ranked public research university situated in the heart of Atlanta, a diverse and vibrant city with multiple universities. Midtown Atlanta, where Georgia Tech is located, has been recognized as one of the 2016 Great Neighborhoods by the American Planning Association due to its liveliness, walkability, and many great cultural and economic strengths. The Institute is a member of the University System of Georgia, the Georgia Research Alliance, and the Association of American Universities. Georgia Tech prides itself on its technology resources, collaborations, high-quality student body, and its commitment to diversity, equity, and inclusion.

Applications will be considered until open positions are filled. For full consideration, applicants are encouraged to submit their applications by December 15, 2019. Applicants are encouraged to clearly identify in their cover letter the area(s) that best describe their research interests. All applications must be submitted online at: https://academicjobsonline.org/ajo/jobs/14715.

More information about the School of Computer Science is available at: http://scs.gatech.edu/.

Henry Samueli School of Engineering and Applied Science

University of California, Los Angeles (UCLA)
Tenure-Track or Tenured Faculty Position

The Electrical and Computer Engineering Department in the Henry Samueli School of Engineering and Applied Science at the University of California, Los Angeles (UCLA) is accepting applications for faculty positions. Our primary focus is on tenure-track assistant professors, however distinguished senior-level applicants will also be considered. The Department seeks candidates with a PhD in a related discipline. Salary is commensurate with education and experience.

The Department is seeking outstanding candidates with the potential for exceptional, original, and innovative research, excellence in teaching, and also a clear commitment to enhancing the diversity of the faculty, graduate student population, and of the majors in Electrical and Computer Engineering. Experience in mentoring women and minorities in STEM fields is desired. The Department is interested in all areas of research traditionally associated with Electrical and Computer Engineering as well as areas involving extra-departmental collaborations with the Institute for the Risk Sciences and the School of Medicine.

However, we are particularly interested in attracting applicants in the following broadly defined areas:

- ▸ Computer Architecture and Experimental Embedded Systems, especially with a focus on Machine Learning, Security or Privacy.
- ▸ Foundations of Autonomy, including related areas such as Control, Optimization, Perception, and Cyber-Physical Systems.
- ▸ Composite Devices and Materials, including Material Growth and Fabrication Techniques leading to Novel Electronic, Magnetic, Photonic, and/or Quantum Devices, Systems, and Architectures.
- ▸ Computational Medicine, particularly Computational Genomics, Clinical Machine Learning, Computer Vision applied to Medical Imaging and other areas which span Engineering and Medicine.

Applications will be reviewed starting November 1, 2019 until the positions are filled, and therefore for full consideration, please apply before this date.

The University of California is an Equal Opportunity/Affirmative Action Employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability, age or protected veteran status. For the complete University of California nondiscrimination and affirmative action policy, see: UC Nondiscrimination & Affirmative Action Policy.

Please apply at https://recruit.apo.ucla.edu/JPF04749.

VIRGINIA TECH™

FACULTY POSITIONS

Department of Computer Science

The Department of Computer Science at Virginia Tech is growing rapidly. Thanks to substantial multi-year investments from the Commonwealth of Virginia combined with significant infrastructure investments by Virginia Tech, we anticipate hiring multiple faculty members at all ranks and in all areas for the next several years. The majority of new positions will be at our main campus in Blacksburg, VA. We also seek outstanding candidates for our program in Northern Virginia, which is rapidly expanding due to Virginia Tech's exciting new **Innovation Campus, (vt.edu/innovationcampus)** in Alexandria, VA, where computer science academic and research programs will play a central role.

We seek candidates at all ranks and in all areas of computer science. The positions offer competitive packages and resources to enable success. Candidates with core research interests in AI/ML, NLP, computer systems, human-computer interaction, cybersecurity, blockchain systems, high-performance computing, computational science, computational biology and bioinformatics, and quantum computing are especially encouraged to apply. Successful candidates will have the opportunity to leverage the department's highly-focused faculty development and mentoring program, as well as numerous successful collaborations with government, national labs, and industry partners.

Candidates for all positions must have a Ph.D. in computer science or a related field at the time of appointment and a rank-appropriate record of scholarship and collaboration in computing research. Successful candidates should give evidence of commitment to issues of diversity in the campus community. Virginia Tech is committed to building a culturally diverse faculty and strongly encourages applications from traditionally underrepresented communities. Tenured and tenure-track faculty will be expected to teach graduate and undergraduate courses, mentor graduate students, and develop a sustainable research group that is internationally recognized for excellence. The positions require occasional travel to professional meetings. Selected candidates must pass a criminal background check prior to employment.

The department currently has 52 faculty members, including 46 tenured or tenure-track faculty, 14 early career awardees, and numerous recipients of faculty awards from IBM, Intel, AMD, Microsoft, Google, Facebook, and others. CS faculty also provide leadership in several interdisciplinary research centers, such as the **Center for Human-Computer Interaction, (hci.vt.edu)** and the **Discovery Analytics Center, (dac.cs.vt.edu).** The department is home to over 1,000 undergraduate majors and 300 graduate students, with university commitments to grow all programs significantly. The department is in the College of Engineering, whose undergraduate program ranks 13th and graduate program ranks 31st among all U.S. engineering schools (*USN&WR*). Virginia Tech's main campus is located in Blacksburg, VA, in a region consistently ranked among the country's best places to live. Our growing program in Northern Virginia offers graduate education and research with one-of-a-kind proximity to government and industry partners.

Applications must be submitted online to **jobs.vt.edu** for position job 510994. Candidates with a clear campus preference (Blacksburg or Northern Virginia) should indicate this in their cover letter. Inquiries should be directed to Dr. Ali R. Butt, search committee chair, at **facdev@cs.vt.edu**.

Virginia Tech is an equal opportunity/affirmative action institution.
A criminal background check is the condition of employment with Virginia Tech.

Illinois Institute of Technology

Tenure-Track/Tenured Positions in Computer Science

The Department of Computer Science at the Illinois Institute of Technology invites applications for multiple tenure-track/tenured faculty positions at all ranks, appointments to start in Fall 2020.

Applicants must have a Ph.D. in computer science or a closely related field, demonstrated excellence in research, a record of attracting external research funding appropriate to their rank, and a strong commitment to teaching. We seek outstanding candidates in all areas of computer science; candidates in cybersecurity, data science, artificial intelligence, parallel and distributed systems, and programming languages are especially encouraged to apply.

The Department of Computer Science at the Illinois Institute of Technology offers Bachelors, Masters, and Ph.D. degrees in Computer Science, as well as Bachelors and Masters degrees in Artificial Intelligence, a Masters degree in Cybersecurity, and interdisciplinary Masters degrees in Data Science and in Computational Decision Science and Operations Research. The department is in a significant growth phase, with multiple faculty hires per year expected for at least the next few years. It is also launching diverse new interdisciplinary research and education programs, and has strong growing partnerships with Chicago's burgeoning tech community. Illinois Institute of Technology, a private, technology-focused research university, is located just 10 minutes from downtown Chicago. The university has recently completed a successful capital campaign that led to the creation of multiple endowed positions, increased scholarship funding, the Center for Active Computational Thinking, and the new Ed Kaplan Family Institute for Innovation and Tech Entrepreneurship. In addition to its rigorous research and education programs, Illinois Tech has a long history of strong partnerships and collaborations with local companies, government labs, and nonprofits; the University Technology Park on campus is home to many startups who benefit from close collaboration with faculty and students.

Review of applications will start on November 1, 2019; applications will be reviewed continually until all available positions are filled. Illinois Institute of Technology is an EEO/AA/Title VI/Title IX/Section 504/ADA/ADEA employer committed to enhancing equity, inclusion and diversity within its community. It actively seeks applications from all individuals regardless of race, color, sex, marital status, religion, creed, national origin, disability, age, military or veteran status, sexual orientation, and/or gender identity and expression. All qualified applicants will receive equal consideration for employment. Applicants should apply online at https://academicjobsonline.org/ajo/jobs/14362.

Pennsylvania State University

Associate Dean for Undergraduate and Graduate Studies (ADUGS)

The College of Information Sciences and Technology (IST) (http://ist.psu.edu) at the Pennsylvania State University in University Park, Pennsylvania invites applications for Associate Dean for Undergraduate and Graduate Studies (ADUGS); although the position is a full-time administrative job, the ADUGS will be a Full or Associate Professor, with tenure in IST. In this regard, candidates for Associate Professor should have a strong track record of research, publication, and funding; those for Full should have a track record of research, publication, and funding that distinguishes them as national or international leaders in their fields.

We seek candidates with a strong record of: graduate and/or undergraduate teaching; graduate and/or undergraduate advising; curriculum development; faculty leadership; scholarship and research. We particularly invite candidates who can contribute to these themes: rapid growth; technology innovation; interdisciplinary connections; undergraduate research; diverse student population. Responsibilities will include: strengthen ties to relevant University Offices; facilitate collaboration within and across the college and university; program assessment and curricular reform; academic integrity; and teaching and research consonant with administrative responsibilities.

Successful candidates must have a Ph.D. or terminal degree in a field relevant to our interdisciplinary faculty (e.g., information and computer science, psychology, sociology) and must pass a background check. To apply: submit basic information via http://psu.jobs/job/90506 and apply via https://academicjobsonline.org/ajo/jobs/14792, sharing your vision for IST education; a CV; and contact information of 4-6 references.

NUS
National University of Singapore

HEAD AND PROFESSOR DEPARTMENT OF COMPUTER SCIENCE

National University of Singapore (NUS) seeks to appoint a distinguished scholar, with outstanding academic and administrative leadership record as Head, Department of Computer Science. The new Head will build upon the strengths and illustrious achievements of the Department and lead the Department in all administrative and academic matters, including strategic planning, research and program development and funding.

The Department of Computer Science, with over 80 faculty members, has a long track record in grooming leaders for the digital economy. The Department currently has approximately S$160 million of funded research from various granting agencies and its internationally recognized faculty members conduct research in the areas of Artificial Intelligence, Database Management, Media, Systems and Networking, Computational Biology, Theoretical Computer Science, Programming Languages and Software Engineering.

Established in 1998, NUS School of Computing has two Departments: Computer Science and Information Systems & Analytics. It has a student population of about 4,400, out of which 800 are graduate students. For more information, please visit: ***www.comp.nus.edu.sg***

NUS is Singapore's flagship university which offers a global approach to education and research with a focus on Asian perspectives and expertise. Its 17 faculties and schools provide a broad-based curriculum underscored by multi-disciplinary courses and cross-faculty enrichment. Its transformative education includes programmes such as student exchange, entrepreneurial internships at 11 NUS Overseas Colleges, as well as double degree, concurrent degree, and joint degree programmes with some of the world's top universities.

NUS is an equal opportunity employer that offers highly competitive salaries and is situated in Singapore, an English-speaking cosmopolitan city that is a melting pot of many cultures, both the east and the west. Singapore offers high-quality education and healthcare at all levels, as well as very low tax rates.

Please send applications and nominations, with detailed curriculum vitae with a list of publications, research and teaching statements together with the names/contacts of at least three references by **15 December 2019** to Distinguished Professor Ooi Beng Chin, Chair of Head Search Committee at the e-mail address *cshodrec@comp.nus.edu.sg*.

Open Rank **Faculty Position** in **Computer Science**

LEARN MORE

https://apply.interfolio.com/67665

Applicant review will begin October 15, 2019 and continue until the position is filled. Please direct inquiries to ADUGSrecruiting@ist.psu.edu.

The Pennsylvania State University is the land grant institution of Pennsylvania. University Park is the largest of Penn State's 24 campuses; undergraduate enrollment is approximately 44K and we offer over 150 graduate programs. Our College has award-winning faculty and state-of-the-art facilities. Both faculty and students are dedicated to collaboration and applying knowledge to make our lives better. University Park is located in State College, Pennsylvania, ranked 3rd safest metropolitan area in the United States by CQ Press, and 8th best college town in the nation by Best College Reviews.

To review the Annual Security Report which contains information about crime statistics and other safety and security matters and policies, please go to https://police.psu.edu/annual-security-reports, which will also explain how to request a paper copy of the Annual Security Report.

Purdue University

Department of Computer Science

Assistant/Associate Professor of Practice Positions in Computer Science

The Department of Computer Science in the College of Science at Purdue University solicits applications for two Professor of Practice positions at the Assistant or Associate Professor level.

Qualifications: Applicants should hold a PhD in computer science or a related field, or a BS degree in computer science or a related discipline and commensurate experience in teaching or industry. Applicants should be committed to excellence in teaching, have the ability to teach a broad range of courses in the undergraduate curriculum, have an enthusiasm for teaching and interaction with students, have an interest in on-line development and delivery of courses, and have an interest in advising student team projects. The positions are non-tenure track faculty positions with multi-year contracts. Professors of Practice faculty are actively involved in departmental activities and have professional development opportunities.

The Department and College: The Department of Computer Science offers a stimulating academic environment with active research programs in most areas of computer science. The department offers undergraduate programs in Computer Science and Data Science, and graduate MS and PhD programs, including a Professional MS in Information Security. For more information, see https://www.cs.purdue.edu.

Computer Science is part of the College of Science, which comprises the computing, physical, and life sciences at Purdue. It is the second-largest college at Purdue with over 350 faculty and more than 6,000 students. The College is pursuing significant new initiatives which complement campus-wide plans, including an Integrative Data Science Initiative. Opportunities for collaboration exist across mathematics, probability, statistics, and the physical and life sciences. Purdue itself is one of the nation's leading land-grant universities, with an enrollment of over 41,000 students primarily focused on STEM subjects.

Application Procedure: Please visit http://www.cs.purdue.edu/hiring to apply. Applications need to include (1) a complete curriculum vitae, (2) a teaching statement that includes the teaching philosophy, interests, and experience, and (3) at least three names of reference. Purdue University's Department of Computer Science is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating their past experiences, current interests or activities, and/or future goals to promote a climate that values diversity and inclusion.

A background check will be required for employment in this position. Review of applications and interviews will begin in October 2019 and will continue until positions are filled. Inquiries can be sent to pop-search@cs.purdue.edu.

Purdue University is an EOE/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

Purdue University

Department of Computer Science

Tenure-Track/Tenured Faculty Positions in Theoretical Computer Science

The Department of Computer Science in the College of Science at Purdue University invites appli-

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

ज्ञानं परमं बलम्

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

BITS Pilani

Announcement for faculty positions in Computer Science

LOOKING FOR: HIGH ACHIEVERS

BITS Pilani, announced by MHRD, Govt. of India to be an **Institute of Eminence (IoE) and ranked within World top 500 subject Ranking in Computer Science by QS,** is looking actively for accomplished researchers in all areas of Computer Science including Data Science, AI, Information Security, Theoretical Computer Science, and High Performance Computing.
The candidates are expected to have predilection for teaching courses in the latest technology.

These faculty positions are open for **Pilani, Goa, Hyderabad and Dubai** campuses of BITS Pilani. An individual with a PhD degree in Computer Science may apply for the positions of Assistant Professor, Associate Professor and Professor. Overseas Citizens of India (OCIs), Persons of Indian Origin (PIOs) and foreign nationals are also encouraged to apply.

BITS Pilani provides its faculty competitive salary and benefits comparable to the best in the country. Startup and contingency funds are also provided to support research.

For detailed information on pay package & benefits and to apply through the online portal visit us at: **www.bits-pilani.ac.in/facultypositions**

Applications received are reviewed periodically for further processing.

innovate achieve lead

cations for two or more tenure-track or tenured positions in theoretical computer science. These appointments will be at the level of Assistant or Associate Professor. The positions are part of a continued expansion in a large-scale hiring effort across key strategic areas in the College of Science.

Qualifications: The Department is interested in candidates whose work in theoretical computer science focuses on the design and analysis of algorithms, quantum computing, randomness in computation, as well as computational science and engineering. Highly qualified applicants in other areas of theoretical computer science will be considered. Applicants should hold a PhD in Computer Science or a related discipline, have demonstrated excellence in research, and have a strong commitment to teaching. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in department and university activities.

The Department and College: The Department of Computer Science offers a stimulating academic environment with active research programs in most areas of computer science. The department offers undergraduate programs in Computer Science and Data Science, and graduate MS and PhD programs, including a Professional MS in Information Security. For more information, see https://www.cs.purdue.edu.

Computer Science is part of the College of Science, which comprises the computing, physical, and life sciences at Purdue. It is the second-largest college at Purdue with over 350 faculty and more than 6,000 students. The College is pursuing significant new initiatives which complement campus-wide plans, including an Integrative Data Science Initiative. Opportunities for collaboration exist across mathematics, probability, statistics, and the physical and life sciences. Purdue itself is one of the nation's leading land-grant universities, with an enrollment of over 41,000 students primarily focused on STEM subjects.

Application Procedure: Please visit http://www.cs.purdue.edu/hiring to apply. Applications need to include (1) a complete curriculum vitae, (2) a statement of research and a statement of teaching, and (3) at least three names of reference. Purdue University's Department of Computer Science is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating their past experiences, current interests or activities, and/or future goals to promote a climate that values diversity and inclusion.

A background check will be required for employment in this position. Review of applications and interviews will begin in November 2019 and will continue until positions are filled. Inquiries can be sent to TA-search@cs.purdue.edu.

Purdue University is an EOE/AA employer.
All individuals, including minorities,
women, individuals with disabilities, and
veterans are encouraged to apply.

Purdue University

Department of Computer Science

Tenure-Track/Tenured Professors in Computer Science - Artificial Intelligence

The Department of Computer Science in the College of Science at Purdue University invites applications for two or more tenure-track or tenured positions in the broad area of artificial intelligence. These appointments will be at the level of Assistant or Associate Professor. The positions are part of a continued expansion in a large-scale hiring effort across key strategic areas in the College of Science.

Qualifications: The Department is broadly interested in candidates from all areas of Artificial Intelligence. To expand and enhance our existing strengths, we are particularly interested in machine learning, natural language processing, human-computer interaction, vision, and reasoning/decision making. Applicants should hold a PhD in Computer Science or a related discipline, have demonstrated excellence in research, and have a strong commitment to teaching. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in department and university activities.

The Department and College: The Department of Computer Science offers a stimulating academic environment with active research programs in most areas of computer science. The department offers undergraduate programs in Computer

Penn Engineering

MULTIPLE FACULTY POSITIONS

Department of Electrical and Systems Engineering

The School of Engineering and Applied Science at the University of Pennsylvania is growing its faculty by 33% over a five year period. As part of this initiative, the **Department of Electrical and Systems Engineering** is engaged in an aggressive, multi-year hiring effort for multiple tenure-track positions at all levels. Candidates must hold a Ph.D. in Electrical Engineering, Computer Engineering, Systems Engineering, or related area. The department seeks individuals with exceptional promise for, or proven record of, research achievement, who will take a position of international leadership in defining their field of study, and excel in undergraduate and graduate education. Leadership in cross-disciplinary and multi-disciplinary collaborations is of particular interest. We are interested in candidates in all areas that enhance our research strengths in:

1. **Nanodevices and nanosystems** (nanoelectronics, MEMS/NEMS, power electronics, nanophotonics, nanomagnetics, quantum devices, integrated devices and systems at nanoscale),
2. **Circuits and computer engineering** (analog, RF, mm-wave, digital circuits, emerging circuit design, computer engineering, IoT, embedded and cyber-physical systems), and
3. **Information and decision systems** (control, optimization, robotics, data science, network science, communications, information theory, signal and image processing).

Prospective candidates in all areas are strongly encouraged to address large-scale societal problems in energy, transportation, health, food and water, economic and financial networks, social networks, critical infrastructure, and national security. We are especially interested in candidates whose interests are aligned with the school's strategic plan, ***https://www.seas.upenn.edu/about/penn-engineering-2020/***

Diversity candidates are strongly encouraged to apply. Interested persons should submit an online application at ***https://www.ese.upenn.edu/faculty-staff/*** and include curriculum vitae, research, teaching, and diversity statements, as well as at least three references. Review of applications will begin on December 1, 2019.

The University of Pennsylvania is an Equal Opportunity Employer.
Minorities/Women/Individuals with Disabilities/Veterans are encouraged to apply.

Australian
National
University

OPEN RANK TENURE TRACK – MULTIPLE COMPUTER SCIENCE FACULTY POSITIONS

Australian National University | Research School of Computer Science

The ANU College of Engineering and Computer Science (CECS) is undergoing significant change and expansion through the Reimagine investment. This substantial 15-year, commitment will transform traditional engineering and computer science disciplines for the 21st century.

This is an exciting time to join our Faculty and be part of a community that prides itself on delivering cutting-edge research and research-led education to develop future leaders, who will find solutions to some of the world's greatest technological and social challenges.

To enquire about these positions please contact Computer Science Director, Professor Tony Hosking, via email to *director.rscs@anu.edu.au*

Come and enjoy the fantastic Australian lifestyle, while working for a world leading University with outstanding staff benefits, including;

• 17% superannuation
• 26 weeks paid parental leave
• 4 weeks paid annual leave
• Exceptional Professional Development opportunities

www.anu.edu.au/jobs

ANU College of Engineering
& Computer Science

Science and Data Science, and graduate MS and PhD programs, including a Professional MS in Information Security. For more information, see https://www.cs.purdue.edu.

Computer Science is part of the College of Science, which comprises the computing, physical, and life sciences at Purdue. It is the second-largest college at Purdue with over 350 faculty and more than 6,000 students. The College is pursuing significant new initiatives which complement campus-wide plans, including an Integrative Data Science Initiative. Opportunities for collaboration exist across mathematics, probability, statistics, and the physical and life sciences. Purdue itself is one of the nation's leading land-grant universities, with an enrollment of over 41,000 students primarily focused on STEM subjects.

Application Procedure: Please visit www.cs.purdue.edu/hiring to apply. Applications need to include (1) a complete curriculum vitae, (2) a statement of research and a statement of teaching, and (3) at least three names of reference. Purdue University's Department of Computer Science is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating their past experiences, current interests or activities, and/or future goals to promote a climate that values diversity and inclusion.

A background check will be required for employment in this position. Review of applications and interviews will begin in November 2019 and will continue until positions are filled. Inquiries can be sent to ai-search@cs.purdue.edu.

Purdue University is an EOE/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

Purdue University

Department of Computer Science

Tenure-Track/Tenured Professors in Computer Science - Systems

The Department of Computer Science in the College of Science at Purdue University invites applications for two or more tenure-track or tenured positions in the broad area of systems. These appointments will be at the level of Assistant or Associate Professor. The positions are part of a continued expansion in a large-scale hiring effort across key strategic areas in the College of Science.

Qualifications: The Department is interested in candidates whose work focuses on database systems, cyber-physical systems, operating systems, networking and distributed systems. Highly qualified applicants in all areas of systems will be considered. Applicants should hold a PhD in Computer Science or a related discipline, have demonstrated excellence in research, and have a strong commitment to teaching. Successful candidates will be expected to conduct research in their fields of expertise, teach courses in computer science, and participate in department and university activities.

The Department and College: The Department of Computer Science offers a stimulating academic environment with active research programs in most areas of computer science. The department offers undergraduate programs in Computer Science and Data Science, and graduate MS and PhD programs, including a Professional MS in Information Security. For more information, see https://www.cs.purdue.edu.

Computer Science is part of the College of Science, which comprises the computing, physical, and life sciences at Purdue. It is the second-largest college at Purdue with over 350 faculty and more than 6,000 students. The College is pursuing significant new initiatives which complement campus-wide plans, including an Integrative Data Science Initiative. Opportunities for collaboration exist across mathematics, probability, statistics, and the physical and life sciences. Purdue itself is one of the nation's leading land-grant universities, with an enrollment of over 41,000 students primarily focused on STEM subjects.

Application Procedure: Please visit www.cs.purdue.edu/hiring to apply. Applications need to include (1) a complete curriculum vitae, (2) a statement of research and a statement of teaching, and (3) at least three names of reference. Purdue University's Department of Computer Science is committed to advancing diversity in all areas of faculty effort, including scholarship, instruction, and engagement. Candidates should address at least one of these areas in their cover letter, indicating their past experiences, current interests or activities, and/

EPFL

Faculty Positions in Computer and Communication Sciences

at the Ecole polytechnique fédérale de Lausanne (EPFL)

The School of Computer and Communication Sciences (IC) at EPFL invites applications for tenure track faculty positions in all areas of computer and communication sciences.

Areas of particular interest are systems, programming languages and verification, unconventional computing, and intelligent systems. In rare cases, senior faculty appointments may be possible.

We seek candidates with an outstanding academic record and a strong commitment to teaching and mentoring students. EPFL offers internationally competitive salaries, generous research support, attractive start-up resources, and outstanding research infrastructure.

Academics in Switzerland enjoy many research funding opportunities, as well as an exceptionally high standard of living. To apply, please follow the application procedure at https://facultyrecruiting.epfl.ch/position/18186243

The following documents are requested in PDF format: cover letter, curriculum vitae including a publication list, brief statements of research and teaching interests, and contact information (name, postal address, and email) of 3 references. Screening will start on **December 1, 2019**. Further questions can be addressed to:

Profs. George Candea and Rüdiger Urbanke
Co-Chairs of the Faculty Recruiting Committee
recruiting.ic@epfl.ch

For additional information on EPFL and IC, please visit: http://www.epfl.ch or http://ic.epfl.ch

EPFL is an equal opportunity employer and family friendly university. It is committed to increasing the diversity of its faculty. It strongly encourages women to apply.

or future goals to promote a climate that values diversity and inclusion.

A background check will be required for employment in this position. Review of applications and interviews will begin in November 2019 and will continue until positions are filled. Inquiries can be sent to systems-search@cs.purdue.edu.

Purdue University is an EOE/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

Purdue University

School of Electrical and Computer Engineering

Assistant or Associate Professor of Computer Engineering

The School of Electrical and Computer Engineering at Purdue University is seeking applications for tenured or tenure-track positions at the Assistant or Associate Professor level in any area of Computer Engineering. We are particularly interested in candidates in computer systems and computer security. All aspects of computer systems will be considered such as computer networks, mobile computing, operating systems, dependability, and embedded systems. Similarly, all aspects of computer security will be considered including data security and privacy, network security, software security, and systems security.

Successful candidates must hold a Ph.D. degree in Electrical and Computer Engineering, Computer Science, or a related discipline. They should demonstrate an excellent potential to build an independent research program at the forefront of their field, and to educate and mentor students. Successful candidates will conduct original research, advise graduate students, teach undergraduate and graduate level courses, and perform service both at the School and University levels.

These positions are part of a continued expansion in a large-scale hiring effort across key strategic areas in the College of Engineering. Purdue Engineering is pursuing significant new growth and initiatives in Computer & Information Systems Engineering within ECE. These are evidenced by recent strategic investments by the college, ECE, and external sponsors in centers such as C-BRIC, PurPL, and CRISP.

The School is an integral part of Purdue's College of Engineering. Purdue Engineering is one of the largest and highest-ranked engineering colleges in the nation (8th for graduate programs and 9th for undergraduate per US News and World Report, 2019) and renowned for top-notch faculty, students, unique research facilities, and a culture of collegiality and excellence. The College goal of Pinnacle of Excellence at Scale is guiding strategic growth in new directions, by investing in people, exciting initiatives, and facilities.

Submit applications online at https://tinyurl.com/purdue-ecesystems2019, including curriculum vitae, teaching and research plans, names of three references, and copies of the two most significant publications. For information/questions regarding applications, contact the Office of Academic Affairs, College of Engineering, at coeacademicaffairs@purdue.edu. Review of applications will begin on September 16, 2019. Applications received after the date will continue to be reviewed until the positions are filled. A background check will be required for employment in this position.

Purdue University is an EOE/AA employer. All individuals, including minorities, women, individuals with disabilities, and veterans are encouraged to apply.

San José State University - San José, California

Assistant/Associate Professor (Tenure-Track)

San José State University - San José, California
POSITION AVAILABILITY
Subject to Budgetary Approval

Specialization: Computer/Software Engineering
Job Opening ID (JOID): 25110
Rank: Assistant/Associate Professor (Tenure-Track)

The Computer Engineering Department at San José State University (SJSU) invites applications for two tenure-track faculty positions at the rank of Assistant or Associate Professor. Areas of particular interest include machine learning and artificial intelligence, virtual and augmented reality, robotics, data mining and big data, cloud computing and virtualization, networking and mobile systems, computer systems architecture, FPGA, and embedded systems, but other areas

上海科技大学
ShanghaiTech University

TENURE-TRACK AND TENURED POSITIONS

ShanghaiTech University invites highly qualified candidates to fill multiple tenure-track/tenured faculty positions as its core founding team in the School of Information Science and Technology (SIST). We seek candidates with exceptional academic records or demonstrated strong potentials in all cutting-edge research areas of information science and technology. They must be fluent in English. English-based overseas academic training or background is highly desired.

ShanghaiTech is founded as a world-class research university for training future generations of scientists, entrepreneurs, and technical leaders. Boasting a new modern campus in Zhangjiang Hightech Park of cosmopolitan Shanghai, ShanghaiTech shall trail-blaze a new education system in China. Besides establishing and maintaining a world-class research profile, faculty candidates are also expected to contribute substantially to both graduate and undergraduate educations.

Academic Disciplines: Candidates in all areas of information science and technology shall be considered. Our recruitment focus includes, but is not limited to: computer architecture, software engineering, database, computer security, VLSI, solid state and nano electronics, RF electronics, information and signal processing, networking, security, computational foundations, big data analytics, data mining, visualization, computer vision, bio-inspired computing systems, power electronics, power systems, machine and motor drive, power management IC as well as inter-disciplinary areas involving information science and technology.

Compensation and Benefits: Salary and startup funds are highly competitive, commensurate with experience and academic accomplishment. We also offer a comprehensive benefit package to employees and eligible dependents, including on-campus housing. All regular ShanghaiTech faculty members will join its new tenure-track system in accordance with international practice for progress evaluation and promotion.

Qualifications:

- Strong research productivity and demonstrated potentials;
- Ph.D. (Electrical Engineering, Computer Engineering, Computer Science, Artificial Intelligence, Financial Engineering, Signal Processing, Operation Research, Applied Math, Statistics or related field);
- A minimum relevant (including PhD) research experience of 4 years.

Applications: Submit (in English, PDF version) a cover letter, a 2-page research plan, a CV plus copies of 3 most significant publications, and names of three referees to: sist@shanghaitech.edu.cn. For more information, visit http://sist.shanghaitech.edu.cn/2017/0426/c2865a23763/page.htm

Deadline: The positions will be open until they are filled by appropriate candidates.

Department of Electrical and Computer Engineering
Graduate School of Engineering and Management
Air Force Institute of Technology (AFIT)
Dayton, Ohio

Faculty Position

The Department of Electrical and Computer Engineering at the Air Force Institute of Technology is seeking applications for a tenured or tenure-track faculty position. All academic ranks will be considered. Applicants must have an earned doctorate in Electrical Engineering, Computer Engineering, Computer Science, or a closely affiliated discipline by the time of their appointment (anticipated 1 September 2020).

We are particularly interested in applicants specializing in one or more of the following areas: autonomy, artificial intelligence / machine learning, navigation with or without GPS, cyber security, and VLSI. Candidates in other areas of specialization are also encouraged to apply. This position requires teaching at the graduate level as well as establishing and sustaining a strong DoD relevant externally funded research program with a sustainable record of related peer-reviewed publications.

The Air Force Institute of Technology (AFIT) is the premier Department of Defense (DoD) institution for graduate education in science, technology, engineering, and management, and has a Carnegie Classification as a High Research Activity Doctoral University. The Department of Electrical and Computer Engineering offers accredited M.S. and Ph.D. degree programs in Electrical Engineering, Computer Engineering, and Computer Science as well as an MS degree program in Cyber Operations.

Applicants must be U.S. citizens. Full details on the position, the department, applicant qualifications, and application procedures can be found at ***http://www.afit.edu/ENG/*** . Review of applications will begin on January 6, 2020. The United States Air Force is an equal opportunity, affirmative action employer.

in computer and software engineering will also be considered. For a complete job description please go to the Interfolio link below.

For full consideration, send a letter of application, curriculum vitae, statements of (1) teaching interests/ philosophy, (2) research plans, (3) specific diversity initiatives, strategies, activities that have been accomplished and/or are planned to advance diversity, equity, and/or inclusion, and at least three original letters of reference with contact information by January 6, 2020 to https://apply.interfolio.com/67145.

Stanford University

Graduate School of Business

Faculty Positions in Operations, Information and Technology

The Operations, Information and Technology (OIT) area at the Graduate School of Business, Stanford University, is seeking qualified applicants for full-time, tenure-track positions, starting September 1, 2020. All ranks and relevant disciplines will be considered. Applicants are considered in all areas of Operations, Information and Technology (OIT), including the management of service and manufacturing systems, supply and transportation networks, information systems/technology, energy systems, and other systems wherein people interact with technology, markets, and the environment. Applicants are expected to have rigorous training in management science, operations research, engineering, computer science, economics, and/or statistical modeling methodologies. Candidates with strong empirical training in economics, behavioral science or computer science are encouraged to apply. The appointed will be expected to do innovative research in the OIT field, to participate in the school's PhD program, and to teach both required and elective courses in the MBA program. Junior applicants should have or expect to complete a PhD by September 1, 2020.

Applicants should submit their applications electronically by visiting the web site http://www.gsb.stanford.edu/recruiting and uploading their curriculum vitae, research papers and publications, and teaching evaluations, if applicable, on that site. Applications will be accepted until November 30, 2019. **For an application to be considered complete, the applicant must submit a CV and job market paper and arrange for three letters of recommendation to be submitted before the application deadline of November 30, 2019.**

The Stanford Graduate School of Business will not conduct interviews at the INFORMS meeting in Seattle, but some OIT faculty members will attend.

Any questions regarding the application process should be sent by email to Faculty_Recruiter@gsb.stanford.edu.

Stanford is an equal employment opportunity and affirmative action employer. All qualified applicants will receive consideration for employment without regard to race, color, religion, sex, sexual orientation, gender identity, national origin, disability, protected veteran status, or any other characteristic protected by law. Stanford welcomes applications from all who would bring additional dimensions to the University's research, teaching and clinical missions.

Swarthmore College

Multiple Faculty Positions in Computer Science

The Department of Computer Science at Swarthmore College invites applications for (1) a tenure-track position at the rank of Assistant Professor and (2) multiple visiting assistant professor positions to begin fall semester 2020. Applicants must have or expect to have a Ph.D. in Computer Science or a related field by the position's start date. All areas of computer science will be considered. We are particularly interested in areas that complement our existing offerings, including compilers, programming languages, high-performance computing, security, algorithms, and theory. The Department also welcomes candidates who conduct interdisciplinary research in the humanities and social sciences.

Swarthmore College is a highly selective liberal arts college, located in the suburbs of Philadelphia, whose mission combines academic rigor with social responsibility. The Computer Science Department currently has nine tenure-track faculty and three visiting faculty. Faculty teach introductory courses as well as advanced courses in their research areas. Our majors and minors are much more diverse than the national averages in CS and we also have 35% female majors. We have grown significantly in both faculty and students

in the last five years. Presently, we are one of the most popular majors at the College and expect to have over 70 Computer Science majors graduating this year (2020).

Qualifications:
Applicants must have a Ph.D. in Computer Science or expected by fall 2020. Applicants strong in any area of computer science will be considered.

Institutional Statement on Teaching Diverse Audiences:
The strongest candidates will be expected to demonstrate a commitment to creative teaching and an active research program that speaks to and motivates undergraduates from diverse backgrounds.

Applicant Instructions:
Applicants should include a cover letter, a curriculum vitae, a research statement, a teaching statement and three letters of recommendation, including at least one letter specifically commenting on teaching. Applications will not be considered until letters of recommendation have been submitted. Please address any questions you may have to Kathy Reinersmann, Computer Science Department at kreiner1@swarthmore.edu.

Applications received by November 15, 2019 for the Tenure Track position will receive full consideration - Apply at https://apply.interfolio.com/67943.

Applications received by January 15, 2020 for the visiting assistant professor position will receive full consideration – Apply at https://apply.interfolio.com/68448.

Review of all applications will continue until the positions are filled.

Swarthmore College actively seeks and welcomes applications from candidates with exceptional qualifications, particularly those with demonstrable commitments to a more inclusive society and world. Swarthmore College is an Equal Opportunity Employer. Women and minorities are encouraged to apply.

Trinity College, Hartford, Connecticut

Assistant Professor of Computer Science

Applications are invited for a tenure-track position in computer science at the rank of Assistant Professor to start in the fall of 2020. Candidates must hold a Ph.D. in computer science at the time of appointment.

We are seeking candidates with teaching and research interests in applied areas associated with data analytics, such as database and information systems, data mining and knowledge discovery, machine learning, and artificial intelligence, but other related areas will also be seriously considered.

Trinity College is a coeducational, independent, nonsectarian liberal arts college located in, and deeply engaged with, Connecticut's capital city of Hartford. Our approximately 2,200 students come from all socioeconomic, racial, religious, and ethnic backgrounds across the United States, and seventeen percent are international. We emphasize excellence in both teaching and research, and our intimate campus provides an ideal setting for interdisciplinary collaboration. Teaching load is four courses per year for the first two years and five courses per year thereafter, with a one-semester leave every four years. We offer a competitive salary and benefits package, plus a start-up expense fund. For information about the Computer Science Department, visit: http://www.cs.trincoll.edu/.

Applicants should submit a curriculum vitae and teaching and research statements and arrange for three letters of reference to be sent to: https://trincoll.peopleadmin.com/postings/2020.

Consideration of applications will begin on December 15, 2019, and continue until the position is filled.

Trinity College is an Equal-Opportunity/Affirmative-Action employer. Women and members of minority groups are encouraged to apply.

The University of Alabama in Huntsville

Assistant Professor

The Department of Computer Science at The University of Alabama in Huntsville (UAH) invites applicants for a tenure-track faculty position at the Assistant Professor level beginning August 2020 to support the gaming and entertainment computing program.

A Ph.D. in computer science or a closely related area is required. The successful candidate will have a strong academic background and be able to secure and perform funded research in areas typical for publication in well-regarded academic conference and journal venues. In addition, the candidate should embrace the opportunity to provide undergraduate education.

The department has a strong commitment to excellence in teaching, research, and service; the candidate should have good communication skills, strong teaching potential, and research accomplishments.

UAH is located in an expanding, high-technology area, in close proximity to Cummings Research Park, the second largest research park in the nation and the fourth largest in the world. Nearby are the NASA Marshall Space Flight Center, the Army's Redstone Arsenal, numerous Fortune 500 and high tech companies. UAH also has an array of research centers, including information technology and cybersecurity. In short, collaborative research opportunities are abundant, and many well-educated and highly technically skilled people are in the area. There is also access to excellent public schools and inexpensive housing.

UAH has an enrollment of approximately 9,900 students. The Computer Science department offers BS, MS, and PhD degrees in Computer Science and contributes to interdisciplinary degrees. Faculty research interests are varied and include cybersecurity, mobile computing, data science, software engineering, visualization, graphics and game computing, multimedia, AI, image processing, pattern recognition, and distributed systems. Recent NSF figures indicate the university ranks 30th in the nation in overall federal research funding in computer science.

Interested parties must submit a detailed resume with references to info@cs.uah.edu or Chair, Search Committee, Department of Computer Science, The University of Alabama in Huntsville, Huntsville, AL 35899. Qualified female and minority candidates are encouraged to apply. Initial review of applicants will begin as they are received and continue until a suitable candidate is found.

The University of Alabama in Huntsville is an affirmative action/equal opportunity employer/minorities/females/veterans/disabled.

Please refer to log number: 20/21-549

University of Central Missouri

Assistant Professor in Computer Science - Multiple Positions

The School of Computer Science and Mathematics at the University of Central Missouri is accepting applications for four tenure-track positions in Computer Science at the rank of Assistant Professor. The appointment will begin August 2020. We are looking for faculty excited by the prospect of shaping our school's future and contributing to its sustained excellence.

The Position: Duties will include teaching undergraduate and graduate courses in computer science and/or cybersecurity and developing new courses depending upon the expertise of the applicant and school needs, conducting research which leads toward peer-reviewed publications and/or externally funded grants, and program accreditation/assessment. Faculty are expected to assist with school and university committee work and service activities and advising majors.

Required Qualifications:

- ▸ Ph.D. in Computer Science by August 2020
- ▸ Research expertise and/or industrial experiences in Cybersecurity, Bioinformatics, Game Development or Software Engineering
- ▸ Demonstrated ability to teach existing courses at the undergraduate and graduate levels
- ▸ Ability to develop a quality research program and secure external funding
- ▸ Commitment to engage in curricular development/assessment at the undergraduate and graduate levels
- ▸ A strong commitment to excellence in teaching, research, and continued professional growth
- ▸ Excellent verbal and written communication skills

The Application Process: To apply online, go to https://jobs.ucmo.edu. Apply to positions #997516, #997517, #998332 or #998446. The following items should be attached: a letter of interest, a curriculum vitae, a teaching and research statement, copies of transcripts, and a list of at least three professional references including their names, addresses, telephone numbers and email addresses. Official transcripts and three letters of recommendation will be requested for candidates invited for on-campus interview.

For more information, contact:

Dr. Songlin Tian, Search Committee Chair
School of Computer Science and
Mathematics
University of Central Missouri
Warrensburg, MO 64093
(660) 543-4930
tian@ucmo.edu

Initial screening of applications begins November 15, 2019 and continues until position is filled. AA/EEO/ADA. Women and minorities are encouraged to apply.

UCM is located in Warrensburg, MO, which is 35 miles southeast of the Kansas City metropolitan area. It is a public comprehensive university with about 12,000 students. The School of Computer Science and Mathematics offers undergraduate and graduate programs in Computer Science, Cybersecurity and Software Engineering with over 1000 students. The undergraduate Computer Science and Cybersecurity programs are accredited by the Computing Accreditation Commission of ABET.

The University of Macau (UM)

State Key Laboratory of Internet of Things for Smart City

Chair/Distinguished/Full/Associate/Assistant Professor

The University of Macau (UM) is the only public comprehensive university in Macao. Leveraging this unique advantage, UM aims to establish itself as a world-class university with regional characteristics. English is its working language. In recent years, UM has seen a significant development in and a rising international recognition for its teaching, research, and community service. It has implemented a unique '4-in-1' education model that integrates discipline-specific education, general education, research and internship education, and community and peer education. Combining this model with the largest residential college system in Asia, UM provides all-round education to students. In addition, it recruits outstanding scholars from around the world to create a multilingual and multicultural learning environment for students. With the development of the Guangdong-Hong Kong-Macao Greater Bay Area, and the new initiatives of the university to boost cutting-edge research and interdisciplinary programmes, UM embraces unprecedented opportunities for development, and offers bright career prospect to professionals in different areas.

The State Key Laboratory of Internet of Things for Smart City (https://skliotsc.um.edu.mo/) invites applications for the position of **Chair/Distinguished/Full/Associate/Assistant Professor**, who will also be a joint faculty member in the Faculty of Science and Technology (http://www.fst.um.edu.mo/), in the following disciplines:

► Chair/Distinguished/Full Professor in Intelligent Sensing and Network Communication (Ref. No.: IOTSC/CDF/ISNC/08/2019)

► Chair/Distinguished/Full Professor in Intelligent Transportation (Ref. No.: IOTSC/CDF/IT/08/2019)

► Associate/Assistant Professor in Intelligent Sensing and Network Communication (Ref. No.: IOTSC/AAP/ISNC/08/2019)

► Associate/Assistant Professor in Urban Big Data and Intelligent Technology (Ref. No.: IOTSC/AAP/BD/08/2019)

► Associate/Assistant Professor in Intelligent Transportation (Ref. No.: IOTSC/AAP/IT/08/2019)

► Associate/Assistant Professor in Urban Public Safety and Disaster Prevention (Ref. No.: IOTSC/AAP/UD/08/2019)

The selected candidate is expected to assume duty in January 2020.

Remuneration

A taxable annual remuneration starting from MOP800,800 (approximately USD98,860) as Assistant/Associate Professor and MOP1,170,400 (approximately USD144,490) as Full/Distinguished/Chair Professor will be commensurate with the successful applicants' academic qualification and relevant professional experience. The current local maximum income tax rate is 12% but is effectively around 5% - 7% after various discretionary exemptions. Apart from competitive remuneration, UM offers a wide range of benefits, such as medical insurance, provident fund, on campus accommodation/housing allowance and other subsidies. Further details on our package are available at: https://www.um.edu.mo/admo/vacancy_faq/.

Application Procedure

Applicants should visit **https://career.admo.um.edu.mo/** for more details and to apply **ONLINE**. Review of applications will commence upon receiving applications and continue until the position is filled. Applicants may consider their applications not successful if they are not invited for an interview within 3 months of application.

Human Resources Section,
Office of Administration
University of Macau, Av. da Universidade,
Taipa, Macau, China
Website: https://career.admo.um.edu.mo/;
Email: vacancy@um.edu.mo
Tel: +853 8822 8574; Fax: +853 8822 2412

The effective position and salary index are subject to the Personnel Statute of the University of Macau in force. The University of Macau reserves the right not to appoint a candidate. Applicants with less qualification and experience can be offered lower positions under special circumstances.

Personal data provided by applicants will be kept confidential and used for recruitment purpose only

Under the equal condition of qualifications and experience, priority will be given to Macao permanent residents

University of Michigan, Ann Arbor

Computer Science and Engineering (CSE)

Multiple Tenure-Track and Teaching Faculty (Lecturer) Positions

Computer Science and Engineering (CSE) at the University of Michigan invites applications for multiple tenure-track and teaching faculty (lecturer) positions. We seek exceptional candidates at all levels in all areas across computer science and computer engineering, with special emphasis on candidates at the early stages of their careers. Qualifications include an outstanding academic record, an awarded or expected doctorate or equivalent in computer science or computer engineering, and a strong commitment to teaching and research. Candidates are expected, through their research, teaching, and/or service, to contribute to the diversity and excellence of the academic community. We also have a targeted search for an endowed professorship in theoretical computer science (the Fischer Chair).

The University of Michigan is one of the world's leading research universities, consisting of highly ranked departments and colleges across engineering, sciences, medicine, law, business, and the arts, with a commitment to interdisciplinary collaboration. CSE is a vibrant and innovative community, with over 70 world-class faculty members, over 300 graduate students, and a large and illustrious network of alumni. Ann Arbor is known as one of the best small cities in the nation. The University of Michigan has a strong dual-career assistance program.

We encourage candidates to apply as soon as possible. Positions remain open until filled and applications can be submitted throughout the year.

For more details on these positions and to apply, please visit https://cse.engin.umich.edu/about/faculty-hiring/.

Michigan Engineering's vision is to be the world's preeminent college of engineering serving the common good. This global outlook, leadership focus, and service commitment permeate our culture. Our vision is supported by our mission and values that, together, provide the framework for all that we do. Information about our vision, mission and values can be found at: **http://strategicvision.engin.umich.edu/**.

The University of Michigan has a storied legacy of commitment to Diversity, Equity and Inclusion (DEI). The Michigan Engineering component of the University's comprehensive, five-year, DEI strategic plan—with updates on our programs and resources dedicated to ensuring a welcoming, fair, and inclusive environment—can be found at: **http://www.engin.umich.edu/college/about/diversity**.

The University of Michigan is a Non-Discriminatory/Affirmative Action Employer.

University of Michigan - Dearborn

Assistant Professors in Computer and Information Science (CIS)

The Department of Computer and Information Science (CIS) (https://umdearborn.edu/cecs/departments/computer-and-information-science) at the University of Michigan - Dearborn (https://umdearborn.edu/) invites applications for two tenure-track Assistant Professor positions. Applicants in the area of software engineering will be considered for the first position, while applicants in all areas of computer science, with preference given to areas related to emerging systems (including IoT, edge/cloud computing, visualization, VR/AR, etc.), will be considered for the second position. The expected starting date is September 1, 2020. Although candidates at the Assistant Professor rank are preferred, exceptional candidates may be considered for the rank of Associate Professor depending upon experience and qualifications. We offer competitive salaries and start-up packages.

The CIS Department offers several B.S. and M.S. degrees, and a Ph.D. degree. The current research areas in the department include artificial intelligence, computational game theory, computer graphics, cybersecurity, data privacy, data science/management, energy-efficient systems, game design, graphical models, machine learning, multimedia, natural language processing, networking, service and cloud computing, software engineering, and health informatics. These areas of research are supported by several established labs and many of these areas are currently funded by federal agencies and industries.

Qualifications:
Qualified candidates must have earned a Ph.D. degree in computer science or a closely related discipline by September 1, 2020. Candidates will be expected to do scholarly and sponsored research, as well as teaching at both the undergraduate and graduate levels.

Applications:
Applicants should send a cover letter, curriculum vitae, statements of teaching and research interests, evidence of teaching performance (if any), and a list of three references through Interfolio at:

http://apply.interfolio.com/68333 for the position in **software engineering**;

http://apply.interfolio.com/68336 for the position in **emerging systems or any other area of computer science**.

Review of applications will begin immediately and continue until suitable candidates are appointed.

The University of Michigan-Dearborn, as an equal opportunity/affirmative action employer.

University of South Carolina
Artificial Intelligence Institute
Multiple Open-Rank Faculty Positions

The Artificial Intelligence (AI) Institute (http://ai.sc.edu) is a new university-wide institute engaged in core AI research, as well as high-impact interdisciplinary research involving AI implementations and applications. It is an outcome of the university's Presidential Excellence Initiative, which seeks to bring national prominence to our college and university through AI research and its economic impact. We seek multiple tenured and tenure-track faculty members at all ranks in core-AI and in interdisciplinary fields at the intersection with engineering disciplines.

▸ Applicant is required to possess a Ph.D. degree in computer science or a closely related field by the beginning date of employment and have a demonstrated superior record of research accomplishments.

▸ The successful applicant is expected to develop internationally recognized, externally-funded research programs that broaden the institution's strengths, leverage interdisciplinary collaborations (http://bit.ly/AIInst), and align with vital cross-cutting research themes (eg. smart & connected communities, healthcare transformations, and agile manufacturing).

Research areas of special interest include:

▸ Human in the loop or knowledge-enhanced AI, deep learning/MMML, NLP, QA/conversational AI, brain-inspired computing;

▸ AI and Big data (incl. sensor, social, health, biological);

▸ AI and computer vision, robotics, CPS, human-computer interaction, autonomous vehicles, etc.

The faculty will have the appointment with the new AI Institute with tenure-track or tenured appointment in CSE (http://cse.sc.edu) or another department in the college (http://cec.sc.edu/). CEC is ranked among top 100 engineering colleges in the nation, and has many NSF CAREER Award recipients (e.g., CSE has 10). Teaching load is very attractive. The AI Institute has exceptional infrastructure and resources including 20,000 sq. ft. space.

Review of applications will begin November 1, 2019 and continue until positions are filled. Expected start date January 1, 2020 or later. All applicants must apply online at http://uscjobs.sc.edu/postings/67450. Qualified candidates must include: (1) letter of intent, (2) curriculum vitae, (3) concise description of research plans, (4) teaching plan, and (5) names and contact information of 3 references for a junior faculty rank and 5 references for a senior faculty rank (references can be contacted later in the process for a senior position). For questions or further information, please contact Dr. Amit Sheth (amit@sc.edu).

The University of South Carolina does not discriminate in educational or employment opportunities on the basis of race, sex, gender, age, color, religion, national origin, disability, sexual orientation, genetics, protected veteran status, pregnancy, childbirth or related medical conditions.

University of Toronto
Assistant Professor, Teaching Stream

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE) at the University of Toronto invites applications for a full-time teaching stream faculty appointment at the rank of Assistant Professor, Teaching Stream, in the general area of Computer Systems and Software. The appointment will commence on July 1, 2020, or shortly thereafter.

Applicants must have a Ph.D. in Electrical and Computer Engineering, or a related field, at the time of appointment or soon after.

The successful candidate will have demonstrated excellence in teaching and pedagogical inquiry, including in the development and delivery of undergraduate courses and laboratories, curriculum development, and supervision of undergraduate design projects. This will be demonstrated by strong communication skills, a compelling statement of teaching submitted as part of the application highlighting areas of interest, awards and accomplishments and teaching philosophy; sample course syllabi and materials; and teaching evaluations, as well as strong letters of reference from referees of high standing endorsing excellent teaching and commitment to excellent pedagogical practices and teaching innovation.

Eligibility and willingness to register as a Professional Engineer in Ontario is highly desirable.

Salary will be commensurate with qualifications and experience.

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto ranks among the best in North America. It attracts outstanding students, has excellent facilities, and is ideally located in the middle of a vibrant, artistic, diverse and cosmopolitan city. Additional information may be found at http://www.ece.utoronto.ca.

Review of applications will begin after October 9, 2019, however, the position will remain open until December 2, 2019.

As part of your online application (https://utoronto.taleo.net/careersection/jobdetail.ftl?job=1903901&lang=en), please include a cover letter, a curriculum vitae, and a teaching dossier including a summary of your previous teaching experience, your teaching philosophy and accomplishments, your future teaching plans and interests, sample course syllabi and materials, and teaching evaluations. Applicants must arrange for three letters of reference, including at least one primarily addressing the candidates teaching, to be sent directly by the referees (on letterhead, signed and scanned), by email to the ECE department at search2019@ece.utoronto.ca. Applications without any reference letters will not be considered; it is your responsibility to make sure your referees send us the letters while the position remains open.

You must submit your application online while the position is open, by following the submission guidelines given at http://uoft.me/how-to-apply. Applications submitted in any other way will not be considered. We recommend combining attached documents into one or two files in PDF/MS Word format. If you have any questions about this position, please contact the ECE department at search2019@ece.utoronto.ca.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from racialized persons / persons of colour, women, Indigenous / Aboriginal People of North America, persons with disabilities, LGBTQ persons, and others who may contribute to the further diversification of ideas.

As part of your application, you will be asked to complete a brief Diversity Survey. This survey is voluntary. Any information directly related to you is confidential and cannot be accessed by search committees or human resources staff. Results will be aggregated for institutional planning purposes. For more information, please see http://uoft.me/UP.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

University of Toronto
Assistant Professor – Tenure Stream

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE) at the University of Toronto invites applications for up to three full-time tenure stream faculty appointments at the rank of Assistant Professor. The appointments will commence on July 1, 2020, or shortly thereafter.

Within the general field of electrical and computer engineering, we seek applications from candidates with expertise in one or more of the following strategic research areas: 1. Computer Systems and Software; 2. Electrical Power Systems; 3. Systems Control, including but not limited to autonomous and robotic systems.

Applicants must have a Ph.D. in Electrical and Computer Engineering, or a related field, at the time of appointment or soon after.

Successful candidates will be expected to initiate and lead an outstanding, innovative, independent, competitive, and externally funded research program of international calibre, and to teach at both the undergraduate and graduate levels. Candidates must have demonstrated excellence in research and teaching. Excellence in research is evidenced primarily by publications or forthcoming publications in leading journals or conferences in the field, presentations at significant conferences, awards and accolades, and strong endorsements by referees of high international standing. Evidence of excellence in teaching will be demonstrated by strong communication skills; a compelling statement of teaching submitted as part of the application highlighting areas of interest, awards and accomplishments, and teaching philosophy; sample course syllabi and materials; and teaching evaluations, as well as strong letters of recommendation.

Eligibility and willingness to register as a Professional Engineer in Ontario is highly desirable.

Salary will be commensurate with qualifications and experience.

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto ranks among the best in North America. It attracts outstanding students, has excellent facilities, and is ideally located in the middle of a vibrant, artistic, diverse and cosmopolitan city. Additional information may be found at http://www.ece.utoronto.ca.

Review of applications will begin after October 9, 2019, however, the position will remain open until December 2, 2019.

As part of your online application (https://utoronto.taleo.net/careersection/jobdetail.ftl?job=1903700&lang=en), please include a cover letter, a curriculum vitae, a summary of your previous research and future research plans, up to three representative publications, as well as a teaching dossier including a statement of teaching experience and interests, your teaching philosophy and accomplishments, and teaching evaluations. Applicants must arrange for three letters of reference to be sent directly by the referees (on letterhead, signed and scanned), by email to the ECE department at search2019@ece.utoronto.ca. Applications without any reference letters will not be considered; it is your responsibility to make sure your referees send us the letters while the position remains open.

You must submit your application online while the position is open, by following the submission guidelines given at http://uoft.me/how-to-apply. Applications submitted in any other way will not be considered. We recommend combining attached documents into one or two files in PDF/MS Word format. If you have any questions about this position, please contact the ECE department at search2019@ece.utoronto.ca.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from racialized persons / persons of colour, women, Indigenous / Aboriginal People of North America, persons with disabilities, LGBTQ persons, and others who may contribute to the further diversification of ideas.

As part of your application, you will be asked to complete a brief Diversity Survey. This survey is voluntary. Any information directly related to you is confidential and cannot be accessed by search committees or human resources staff. Results will be aggregated for institutional planning purposes. For more information, please see http://uoft.me/UP.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

University of Toronto

Associate Professor – Tenure Stream

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering (ECE) at the University of Toronto invites applications for up to three full-time tenure stream faculty appointment at the rank of Associate Professor. The appointments will commence on July 1, 2020, or shortly thereafter.

Within the general field of electrical and computer engineering, we seek applications from candidates with expertise in one or more of the following strategic research areas: 1. Computer Systems and Software; 2. Electrical Power Systems; 3. Systems Control, including but not limited to autonomous and robotic systems.

Applicants must have a Ph.D. in Electrical and Computer Engineering, or a related field, and have at least five years of academic or relevant industrial experience.

Successful candidates will be expected to maintain and lead an outstanding, independent, competitive, innovative, and externally funded research program of international calibre, and to teach at both the undergraduate and graduate levels. Candidates must have a demonstrated exceptional record of excellence in research and teaching. Excellence in research is evidenced primarily by sustained and impactful publications in leading journals or conferences in the field, distinguished awards and accolades, presentations at significant conferences and an established high profile in the field with strong endorsements by referees of high international standing. Evidence of excellence in teaching will be demonstrated by excellent communication skills, a compelling statement of teaching submitted as part of the application highlighting areas of interest, awards and accomplishments, and teaching philosophy; sample course syllabi and materials; and teaching evaluations, as well as strong letters of recommendation.

Eligibility and willingness to register as a Professional Engineer in Ontario is highly desirable.

Salary will be commensurate with qualifications and experience.

The Edward S. Rogers Sr. Department of Electrical and Computer Engineering at the University of Toronto ranks among the best in North America. It attracts outstanding students, has excellent facilities, and is ideally located in the middle of a vibrant, artistic, diverse and cosmopolitan city. Additional information may be found at http://www.ece.utoronto.ca.

Review of applications will begin after October 9, 2019, however, the position will remain open until December 2, 2019.

As part of your online application (https://utoronto.taleo.net/careersection/jobdetail.ftl?job=1903708&lang=en), please include a cover letter, a curriculum vitae, a summary of your previous research and future research plans, up to three representative publications, as well as a teaching dossier including a statement of teaching experience and interests, your teaching philosophy and accomplishments, and teaching evaluations. Applicants must arrange for three letters of reference to be sent directly by the referees (on letterhead, signed and scanned), by email to the ECE department at search2019@ece.utoronto.ca. Applications without any reference letters will not be considered; it is your responsibility to make sure your referees send us the letters while the position remains open.

You must submit your application online while the position is open, by following the submission guidelines given at http://uoft.me/how-to-apply. Applications submitted in any other way will not be considered. We recommend combining attached documents into one or two files in PDF/MS Word format. If you have any questions about this position, please contact the ECE department at search2019@ece.utoronto.ca.

The University of Toronto is strongly committed to diversity within its community and especially welcomes applications from racialized persons / persons of colour, women, Indigenous / Aboriginal People of North America, persons with disabilities, LGBTQ persons, and others who may contribute to the further diversification of ideas.

As part of your application, you will be asked to complete a brief Diversity Survey. This survey is voluntary. Any information directly related to you is confidential and cannot be accessed by search committees or human resources staff. Results will be aggregated for institutional planning purposes. For more information, please see http://uoft.me/UP.

All qualified candidates are encouraged to apply; however, Canadians and permanent residents will be given priority.

US Air Force Academy

Assistant Professor

The Department of Computer and Cyber Sciences at the US Air Force Academy seeks to fill a faculty position at the Assistant Professor level. Exceptionally qualified candidates at upper ranks will also be considered. The department is particularly interested in candidates with a background in cybersecurity, but all candidates with a passion for teaching computer science are encouraged to apply.

The Academy is a national service institution, charged with producing leaders of character for the US Air Force. Faculty members are expected to exemplify the highest ideals of professionalism and integrity. The Academy is located in Colorado Springs, an area known for its natural beauty and quality of life. The United States Air Force Academy values the benefits of diversity among the faculty to include a variety of educational backgrounds, professional and life experiences.

For information on how to apply, go to https://www.usajobs.gov and search with the keyword 545526600. US citizenship is required. Candidates with specific questions can contact Dr. Barry Fagin at barry.fagin@usafa.edu.

Worcester Polytechnic Institute

Open Rank Professor - Data Science and Assistant Professor

Looking for faculty colleagues who engage deeply in both teaching and impactful research within a curriculum that embraces project-based learning? Consider joining WPI.

The rapidly growing Data Science program at WPI, one of the first in the world to offer a PhD degree in data science, anticipates hiring full-time tenure-track faculty starting Fall 2020 to strengthen this strategic interdisciplinary area. Outstanding candidates in any area related to Data Science will receive full consideration, including Computer Science, Statistics, or Mathematical Sciences,

The deadline for applications is December 10, 2019. Applications will be considered after that date until the position is filled.

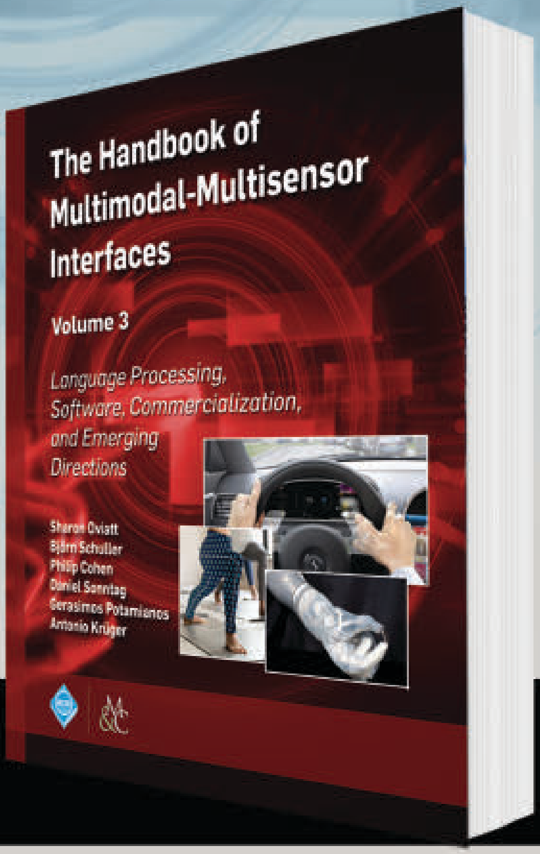
WPI is an Equal Opportunity Employer.

For a detailed position description and to apply, visit: https://apptrkr.com/1626209.

The Handbook of Multimodal-Multisensor Interfaces, Volume 3

Language Processing, Software, Commercialization, and Emerging Directions

This third volume of **The Handbook of Multimodal-Multisensor Interfaces** focuses on state-of-the-art multimodal language and dialogue processing, including semantic integration of modalities. The development of increasingly expressive embodied agents and robots has become an active test-bed for coordinating multimodal dialogue input and output, including processing of language and nonverbal communication. In addition, major application areas are featured for commercializing multimodal-multisensor systems, including automotive, robotic, manufacturing, machine translation, banking, communications, and others. These systems rely heavily on software tools, data resources, and international standards to facilitate their development. For insights into the future, emerging multimodal-multisensor technology trends are highlighted for medicine, robotics, interaction with smart spaces, and similar topics. Finally, this volume discusses the societal impact of more widespread adoption of these systems, such as privacy risks and how to mitigate them. The handbook chapters provide a number of walk-through examples of system design and processing, information on practical resources for developing and evaluating new systems, and terminology and tutorial support for mastering this emerging field. In the final section of this volume, experts exchange views on a timely and controversial challenge topic, and how they believe multimodal-multisensor interfaces need to be equipped to most effectively advance human performance during the next decade.



Edited by Sharon Oviatt et al

ISBN: 978-1-970001-72-3
DOI: 10.1145/3233795
http://books.acm.org
http://store.morganclaypool.com/acm

ACM BOOKS
Collection I

[CONTINUED FROM P. 160] Finally, and most important, EEG sensing, passive for the live audience, active for me and the VR/AR audience and for any live audience members who "enrolled" their sensory implants. Some of the audience know what's happening: manipulation of their emotions, via my voice, which they are steering with their emotions. They don't care, though. They're there to feel something, and the more they give in, give up, give, the better.

But is it Really Music?

Maybe not, or maybe more than ever in history. All I know is that it sure beats the heck out of what I was doing for a living. I have three degrees: Music, AI, and CyberEthics, yet there I was, grinding BlockCoin in the VGame industry. eSports, my ass; just cramps in my hands and fingers, wearing out my tendons and many VR controllers, just to earn a few μB₵s per minute. I have a fairly good voice, and grew up singing, so becoming an AVeC was a far preferable career choice.

The surgeries for the implants didn't hurt much and took only about six weeks to heal. That also gave me time to learn some AI-composed pop songs that MusiCorp™ fed to me. Having all that stuff installed wasn't so hard, but ripping it out would be a far different matter; it could destroy my voice, leaving me unable to speak normally ever again. Also, AVeCs become dependent on it, the feel of the extra hardware, but more importantly, that direct emotional feedback from other humans. There's more than one story of an AVeC having their "rig" removed, and committing suicide within a couple of months, from the pain of the lost (unnatural) human connection.

No, thanks, I'll keep my AVI. Sure, I probably only have another year or two before the next ASIStar replaces me. but after that I can still be an (in-humanly) effective salesman, or politician.[b] Having one of the most influential voices in history is worth a lot, and it won't much matter what I'm saying: it's how I will say it. I'm engineered to connect, to persuade.

Some of the audience knows what's happening: manipulation of their emotions, via my voice, which they are steering with their emotions.

There aren't many ASIngers; the market can only support a few at a time. Certainly not as many as the castrati (my bio-altered singer ancestors) in their heyday, or the robot drummers that were briefly a craze during the last-gasp days of meth-metal. The socioeconomics of all this is, of course, quite bizarre. Just as SnapGram photo filters caused an epidemic of face and body dysmorphia, so did CyberTune, RoboDyne, and other voice perfection technologies create a rash of personal vocal dissatisfaction. People felt hopeless to ever try to sing. They'd never be any good at it, not like those huge AVRTube stars like Gr3tch@n, Cheetθh, and k!dCRAP.

Voice perfection tech meant anyone performing live had to be better than the best singers from before. Direct emotion manipulation was a fairly reliable means to that end. As ASI tech spread, the public quickly grew tired of their AVRTube experiences, and tired of the pop stars that lived there. AR/VR Video channels soon degenerated back to spectacular sports wipe-outs, puppies, kitties, hedgehogs, and similar content. But the music and music personalities left.

So as AR/VR Music Video collapsed, there was a huge uptick in live+VR music concerts, and that rocked the music industry (again). Revenues shifted to per-minute billing for live concerts, with venue attendees paying a slightly lower rate than the higher-fidelity AR/VR network audience. Of course, ASI-style tech found other uses, notably for prostitution, but we won't go into any detail on that. Some guitar players tried hacking ENM (EmotoNeuroMuscular) interfaces into their arms and hands, with interesting results, but not all good. One poor guy put surplus leg muscle actuators in his hands, and was quite amazing, until two of his fingers tore off, flying into the audience during a particularly enthusiastic guitar shred. That was really funn …

Hey!! Pay attention! PAIN/Itch …

I've caught the eye of one particular girl on the front row. Actually, she's caught my focus. I can't do that. Any one-on-one connection messes up the audience biometrics. There's pain, and lots of that itch … The bad itch. I need to climb back into the ASI furrow and do my job.

OK, a little better now …

But not for the audience. My connection with them is now broken by my distraction with that front-row girl. They're not responding correctly. They're jealous, envious. Some are attracted to her. Like that fateful Courtney Cox and Springsteen incident, the audience is now emoting at and with her, not me. I feel that strongly. ASIAI is unhappy. The audience is unhappy.

Oh, no!

Most important, MusiCorp™ is unhappy. Two RoBouncers have picked up the front-row girl and are "ushering" (carrying) her out of the concert venue. Within two seconds, I feel sideways motion. The AI concert manager is rotating the stage to reveal the next "act" early. My voice fades and the new star's voice replaces mine, my backing track morphing into hers.

OK then. My ASInging career is over. Much more quickly than I rose to cyber singer "stardom," I have fallen, and will never rise again.

But … that girl was really cute. Maybe I can duck out the back door into the alley and find her outside on the street (that is, if the RoBouncers haven't whisked her off to a new acting career).

[b] There were attempts to restrict proliferation of bio-assisted persuasion technology, especially its use by politicians. As expected, "forces" were too strong for any meaningful anti-ASI legislation to succeed.

***P-Ray** is the creative/artistic moniker of **Perry R. Cook**, who is professor emeritus of computer science (also music) at Princeton University. Cook is advisor and IP Strategist to social music company Smule, and co-founder of online arts education company Kadenze.

From the intersection of computational science and technological speculation, with boundaries limited only by our ability to imagine what could be.

DOI:10.1145/3363232

P-Ray*

Future Tense
Cantando con la Corrente (Singing with Current)

An augmented singer gets some unexpected feedback from his audience.

AHH … THERE IT IS …

That familiar warm burn, actually more of a sweet pain+itch, guiding me, into the groove, into …

The Flow …

Just go with it. Sing the song. Don't worry about the lyrics. They don't matter much anyway.

Emotion => Affect => Influence.

The slight scoops into certain notes. A touch of vocal fry at the ends of key phrases. Correct pitch, but not that annoying CyberTune™. Just the right amount of breathiness at every instant. Perfect or, actually, maximally influential prosody. My voice, but not completely in my control. I sing the song, sort of. The result: my deep connection with listeners, and theirs with me…

ASIBOV[a] takes care of all that

Beginning a show, from the first song, the warm itch is strong, as ASI helps me do the right things. Bio-actuators ad/abduct my cricothyroid, raise and lower my larynx, flex/pulse my diaphragm, agonist and antagonist, tensing and relaxing all the important parts of my vocal mechanics into just the right places, at just the right times, to create an "optimal" performance. What I can't do physically, ASI takes care of via real-time DSP audio effects. I wear a headset mic anyway, and the audience is far enough away so they hear and feel only perfect, emotional …

IMAGE BY TANDEM BRANDING

Connection …

As I let it happen, I feel it, or rather, I don't feel the itch any more. I am doing the right things. My voice does what ASI wants it to, so the bio-actuators don't have to work so hard to steer me. The differences between what I'm singing and what the audience hears grow ever smaller. The audience yielded long before I did; the AI and DSP took care of that. We all find the flow, in the song, signals, and sensations. Neural nets of silicon and tissue, synchronizing. Layers of machine intelligence grind on bio-emotion signals gathered from the audience: their smiles, open/closed eyes, eye-blinks and rates, breathing rates, body poses, and motions. Also infrared blush detection, hi-definition pupil and iris analysis, even small changes in the levels of CO_2, N_2, O_2, H_2O, and methane in the room.

[CONTINUED ON P. 159]

[a] ASIBOV = Audience+Singer(Speaker) Influenced Bio-feedback Optimized Voice (ASI for short). Invented in 2023 by J.R. Coupling at ARML (Augmented Reality Music Labs), ASIBOV uses analysis of emotional signals gathered in real time from an audience to modify the voice and vocal processing of a singer or speaker. Voice parameters are automatically adjusted for optimized emotional effect. "ASIngers" are also called AVeCs.

‹Programming› 2020

4th International Conference on the Art, Science, and Engineering of Programming

‹Programming› is a conference focused on everything related with programming, including its practice and experience. After Brussels, Nice, and Genova, this year's edition will be hosted by the University of Porto, in Porto, Portugal, a charming city that will embrace you as soon as you arrive!

The program will provide unique opportunities to share knowledge on programming, with keynotes, research papers, workshops, posters, demos, and events with the local academy and industry, in informal and playful settings around the city, for a me-mo-ra-ble experience!

March 23–26, 2020
Porto, Portugal

General Chair › Ademar Aguiar, University of Porto
Program Chair › Stefan Marr, University of Kent
Workshops Chairs › Shigeru Chiba, The University of Tokyo; Elisa Gonzalez Boix, Vrije Universiteit Brussel

Program Committee › Craig Anslow, Edd Barrett, Nicolás Cardozo, Luke Church, Coen De Roover, Erik Ernst, Jun Kato, Jonathan Edwards, Matthew Flatt, Stephen Kell, Diego Garbervetsky, Jeremy Gibbons, Felienne Hermans, Hidehiko Masuhara, Gordana Raki, Guido Salvaneschi, Francisco Sant'Anna, Christophe Scholliers, Friedrich Steimann, Michael Van De Vanter, Didier Verna

https://2020.programming-conference.org

acm In-Cooperation · SIGSOFT · AOSA · U.PORTO FEUP Faculdade de Engenharia Universidade do Porto · Porto.

ICWS
Founded in 2003

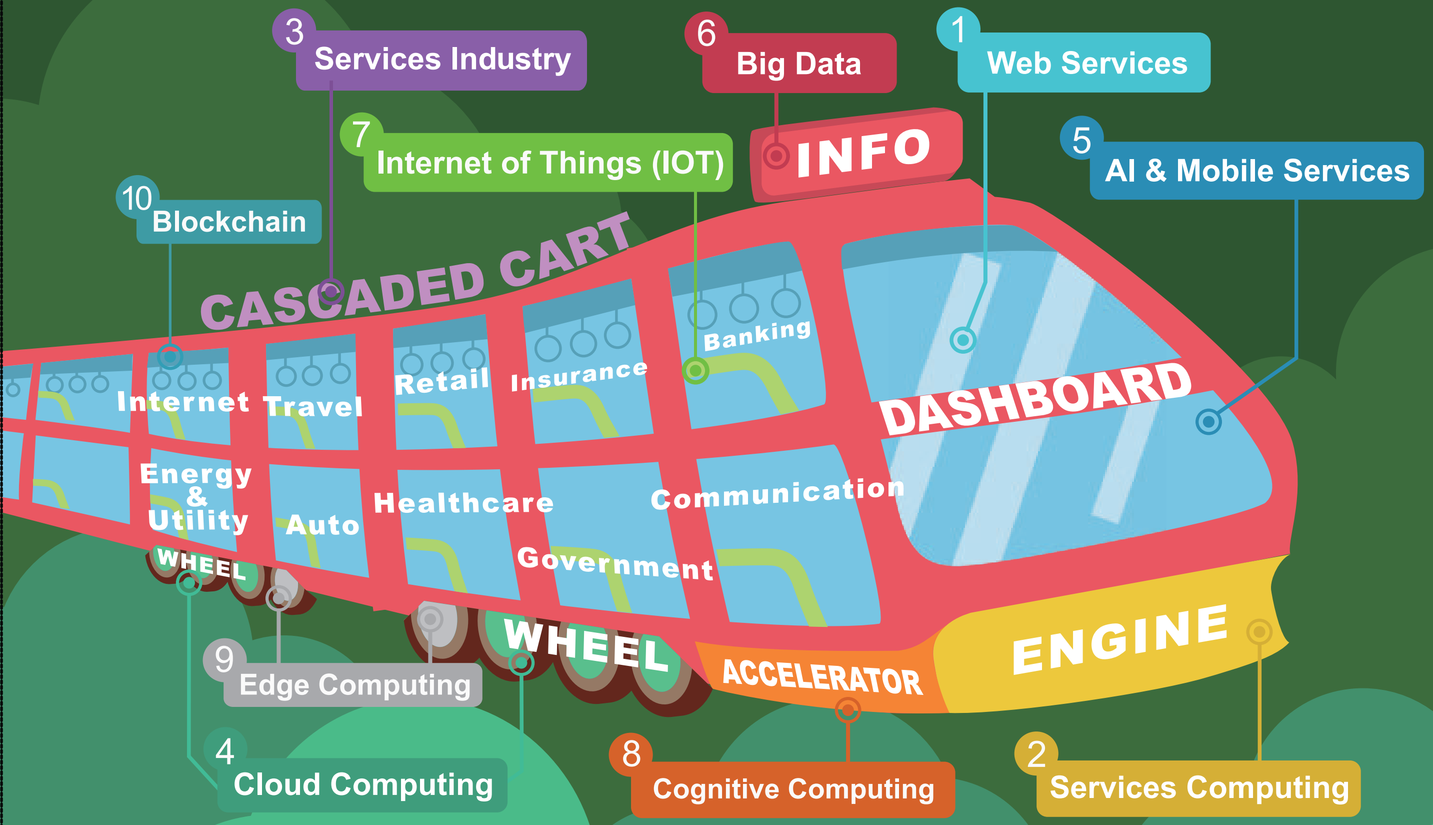
2020 5G for Services Era

JUNE22-26, HONOLULU,HAWAII,USA

CELEBRATING THE 18th GATHERING OF ICWS

SCF
SERVICES
CONFERENCE
FEDERATION

1. 2020 International Conference on Web Services **(ICWS 2020)**
2. 2020 International Conference on Services Computing **(SCC 2020)**
3. 2020 World Congress on Services **(SERVICES 2020)**
4. 2020 International Conference on Cloud Computing **(CLOUD 2020)**
5. 2020 International Conference on AI & Mobile Services **(AIMS 2020)**
6. 2020 International Conference on Big Data **(BigData 2020)**
7. 2020 International Conference on Internet of Things **(ICIOT 2020)**
8. 2020 International Conference on Cognitive Computing **(ICCC 2020)**
9. 2020 International Conference on Edge Computing **(EDGE 2020)**
10. 2020 International Conference on Blockchain **(ICBC 2020)**



3 Services Industry

6 Big Data

1 Web Services

7 Internet of Things (IOT)

5 AI & Mobile Services

10 Blockchain

9 Edge Computing

4 Cloud Computing

8 Cognitive Computing

2 Services Computing

Submission Deadlines:

Early Submission: 12 / 6 / 2019
Regular Submission: 2 / 5 / 2020

Contact:

confs@servicessociety.org / icws.org

SERVICES SOCIETY

Largest not-for-profits organization (501(c)(3)) dedicated for serving 30,000+ worldwide services computing professionals

ICWS.ORG

The Proceedings of SCF 2020 will be published in the LNCS by Springer. Currently the Abstracting and Indexing services covered by Springer's data feeds for LNCS proceedings include ISI Conference Proceedings Citation Index (included in Web of Science), Engineering Index EI (Compendex and Inspec databases), DBLP, Google Scholar, IO-Port, MathSciNet, Scopus, and ZBlMath.