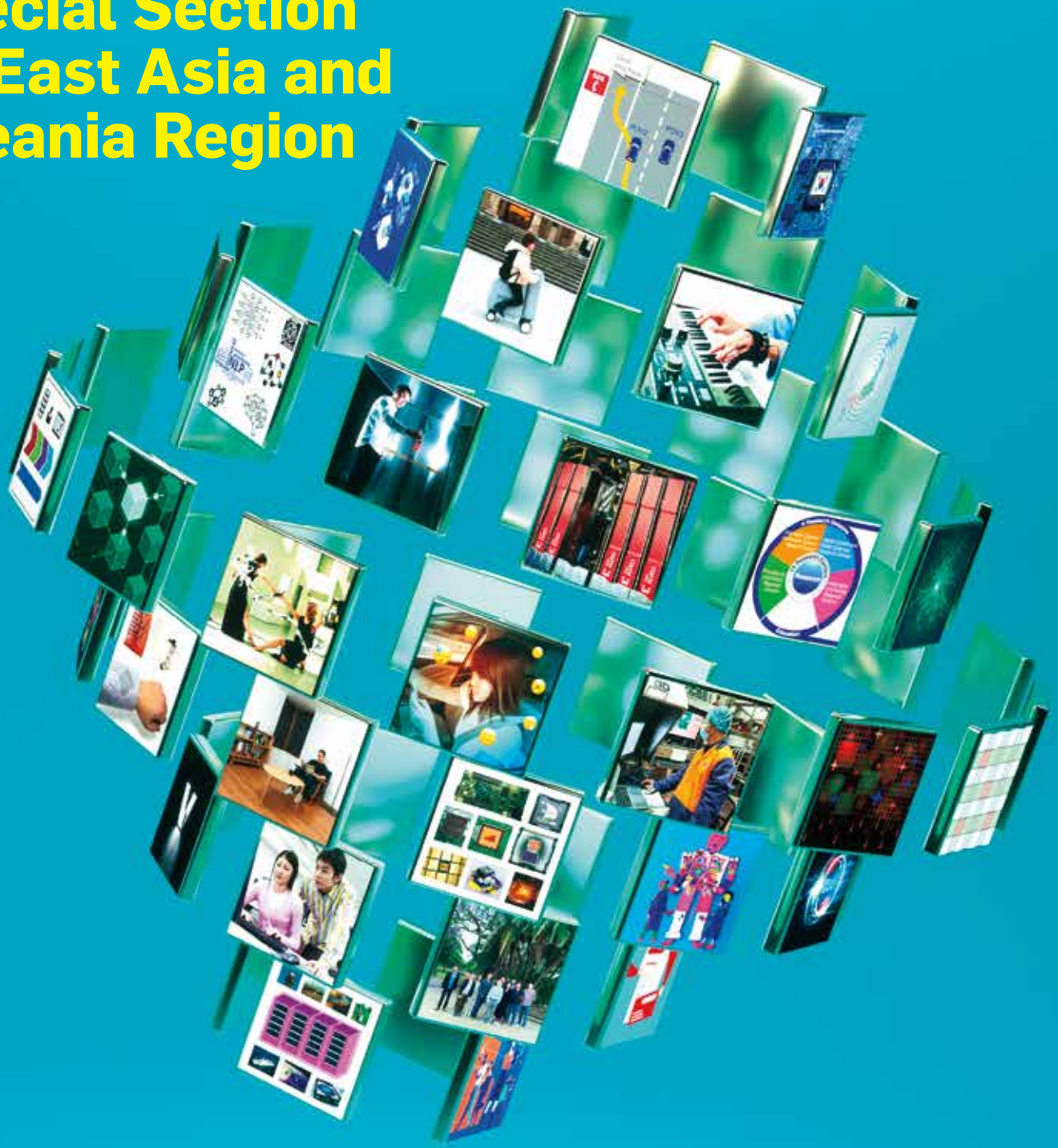
COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

07/2023 VOL.66 NO.07

Special Section on East Asia and Oceania Region

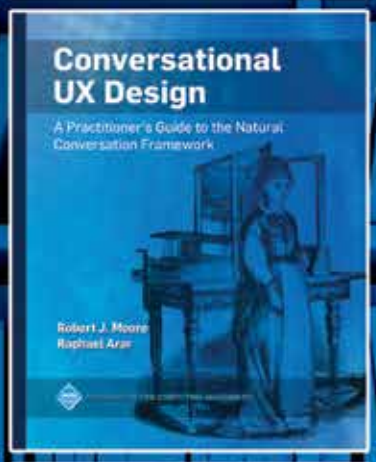


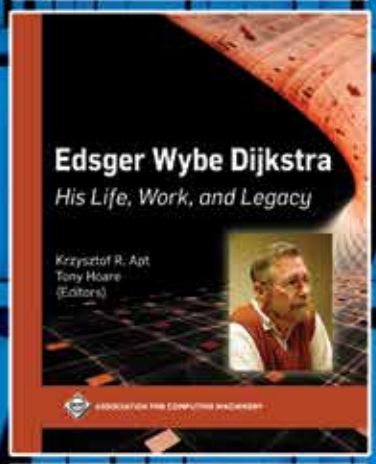
Legal Challenges to Generative AI

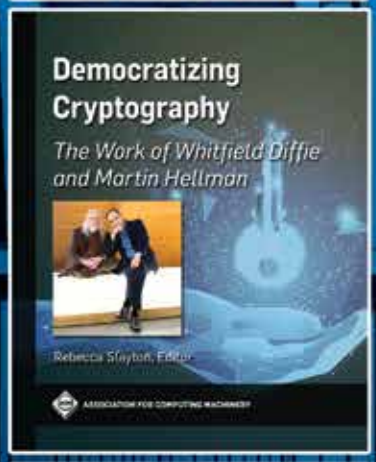
Data Science—A Systematic Treatment

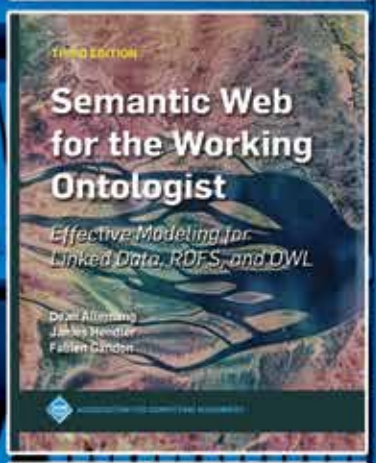
Outsmarting Deepfake Video

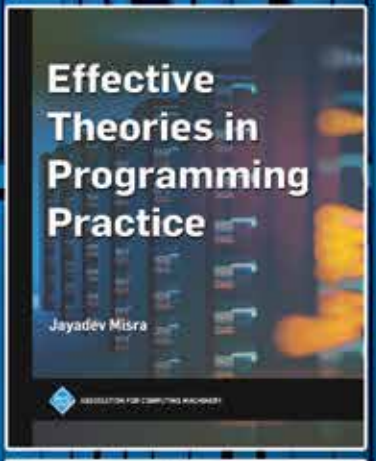
Association for Computing Machinery

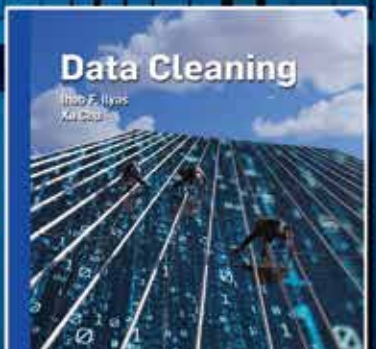












In-Depth. Innovative. Insightful.

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

Title Lists Available for Collection I and Collection II

Collection III began publishing May 2023

ASSOCIATION FOR COMPUTING MACHINERY

For more information, please go to:
http://books.acm.org

1601 Broadway, 10th Floor
New York, NY 10019, USA
212-626-0658
acmbooks-info@acm.org

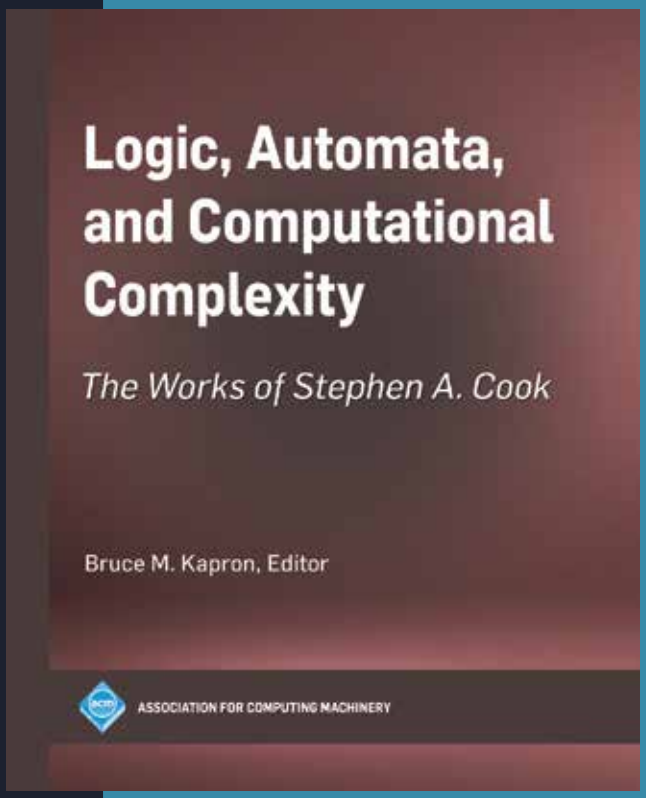
ACM BOOKS

Collection II

Professor Stephen A. Cook is a pioneer of the theory of computational complexity. His work on NP-completeness and the P vs. NP problem remains a central focus of this field. Cook won the 1982 Turing Award for "his advancement of our understanding of the complexity of computation in a significant and profound way."

This volume includes a selection of seminal papers embodying the work that led to this award, exemplifying Cook's synthesis of ideas and techniques from logic and the theory of computation including NP-completeness, proof complexity, bounded arithmetic, and parallel and space-bounded computation. These papers are accompanied by contributed articles by leading researchers in these areas, which convey to a general reader the importance of Cook's ideas and their enduring impact on the research community.

The book also contains biographical material, Cook's Turing Award lecture, and an interview. Together these provide a portrait of Cook as a recognized leader and innovator in mathematics and computer science, as well as a gentle mentor and colleague.



Logic, Automata, and Computational Complexity
The Works of Stephen A. Cook

Bruce M. Kapron, Editor
University of Victoria, BC

ISBN: 9798400707797
DOI: 10.1145/3588287

http://books.acm.org

COMMUNICATIONS OF THE ACM

Departments

Last Byte

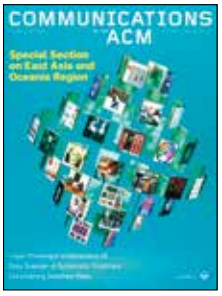
News

Viewpoints

Association for Computing Machinery
Advancing Computing as a Science & Profession

Special Section

Watch the co-organizers discuss this section in the exclusive *Communications* video. https://cacm.acm.org/videos/east-asia-and-oceania-region-2023

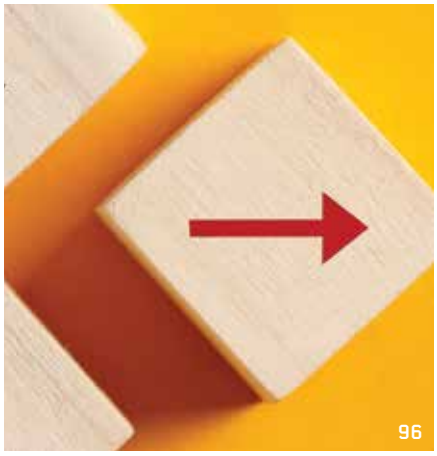


About the Cover: Scenes of technical achievements from the region of East Asia and Oceania. Illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: Photo of Poimo inflatable electric scooter courtesy of University of Tokyo. Remaining images are supplied by authors of East Asia and Oceania Region Special Section articles or are from Shutterstock.com.

IMAGES BY: (L) SPOOKY POOKA AT DEBUT ART; (M) CAGKAN SAYIN; (R) 3D_VICKA

Practice

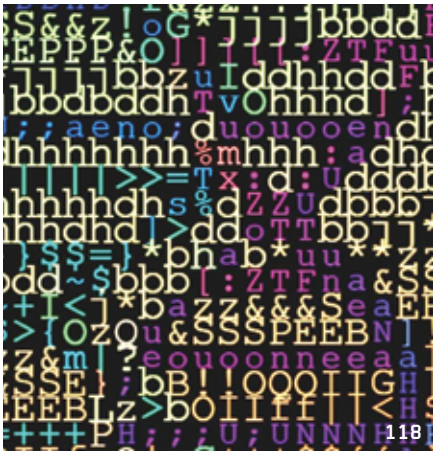


Articles' development led by acmqueue
queue.acm.org

Contributed Articles

Watch the author discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/data-science-systematic-treatment

Review Articles



Research Highlights

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.

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.

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

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

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

DOI:10.1145/3596801

Moshe Y. Vardi

Revisiting ACM's Open-Conference Principle

One of the Presidential Task Forces (PTFs) announced[a] by ACM President Yannis Ioannidis earlier this year focuses on "Open Science." I would like to see that PTF develop a detailed set of guiding principles for conference-site selection. Principles, however, must be translated to concrete decisions in the context of a messy reality. ACM should establish a standing committee on conference-site selection to make such decisions in a deliberate and transparent way.

Let us start with some background. In March 2017, I discussed[b] ACM's Open-Conference Principle, which states: "The open exchange of ideas and the freedom of thought and expression are central to the aims and goals of ACM and its conferences. These aims and goals require an environment that recognizes the inherent worth of every person and group, that fosters dignity, understanding, and mutual respect, and that embraces diversity."

I addressed the tension between this principle and some (then) recent political developments. In 2016–2017, many U.S. state legislators considered or enacted legislation that restricted access to multiuser restrooms, locker rooms, and other sex-segregated facilities based on a certain definition of sex or gender ("bathroom bills"). In response to the North Carolina bathroom bill, ACM's SIGMOD Executive Committee moved its 2017 conference out of North Carolina to a new location. I was critical of that decision.

I argued the bathroom-bill issue was dwarfed by an Executive Order issued by U.S. President Trump in January 2017, which banned nationals of seven Muslim-majority countries from entering the U.S. I asked whether ACM should respond to the travel ban by moving all conferences out of the U.S. I concluded that boycotts may feel right, but they are rarely productive.

Well, the world has become even more complicated since then. In February 2022, Russia launched a full-scale attack against Ukraine, starting the largest military conflict in Europe since World War II. The invasion also caused the largest refugee crisis in Europe. Russian military forces also launched attacks on civilian targets, including hospitals. In response to this brutal aggression, many corporations and associations have severed relationships with Russia.

ACM, however, stayed silent for a while on the issue of conferences in Russia. In March 2022, I started a petition calling on ACM to declare publicly that, while this invasion continues: ACM will not hold or plan to hold any conference in Russia, and ACM will not be affiliated with any conference in Russia. In response to the petition, on March 23, 2022, ACM posted on its website the following text: "On March 3, 2022, ACM's Executive Committee decided not to hold any conferences in Russia while the conflict in the Ukraine and the humanitarian crisis in Europe continue."

Obviously, I had changed my mind on boycotts, at least some boycotts. Why? I considered Russia's conduct to be egregious, so shocking the conscience, that people of good will from around the world must show their opposition to the unprovoked Russian aggression against Ukraine. As I learned, ACM essentially agreed with me.

Recent political developments in the U.S. concerning access to abortion are further testing my no-boycott stance. In 2022, the U.S. Supreme Court overruled *Roe v. Wade*, its own 1973 ruling that the U.S. Constitution generally protects a pregnant individual's liberty to have an abortion. In response to that, many U.S. states have passed laws drastically restricting access to abortion. For example, the state of Florida passed a bill in April 2023 banning all abortions after six weeks of pregnancy. Note that, as of early May, ACM is planning to hold its June 2023 Federated Computing Research Conference (FCRC) in Orlando, FL.

The issue now is the health and safety of conference participants. State abortion bans threaten[c] access to emergency reproductive care[d] by introducing complex legislative restrictions around what counts as a medical emergency. Pregnant individuals who are considering attending FCRC'23 in person may not want to assume this risk. Avoiding holding fully in-person ACM conferences in locations with strict abortion bans is not a boycott. Rather it is a decision that reflects ACM's concern for the well-being of conference participants. In fact, for the same concern, the American Physical Society decided[e] in 2020 to consider police conduct when choosing meeting locations.

This is an important and difficult issue, but ACM must address it. Hence my initial recommendation. C

a https://bit.ly/3FWLtQB
b https://bit.ly/3HPnrrA
c https://bit.ly/44zQyZT
d https://bit.ly/44Ea2MG
e https://bit.ly/3nGfrCp

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

DOI:10.1145/3596212

From UX to Product: An Intentional Career Route

NAME
Hina Elton
BACKGROUND
Born in Leicester, U.K.; now living in Brighton, U.K.
CURRENT JOB TITLE/EMPLOYER
Experience Design Director, Head of Insight and Data, Foolproof, a Zensar company.
EDUCATION
Ph.D. Human-Computer Interaction

Technology has been a part of my life since the age of six. My father was a hardware engineer, fixing all sorts of computers and building various custom hardware solutions for commercial applications. In the late 1980s, he brought home our first family desktop—the IBM 5150. And then came the Apple Macintosh Classic. I was fascinated by the possibilities and tried to learn as much as I could about how these machines worked. None of my friends had computers at home; I felt so lucky.

My mother, a stroke-sufferer since her early 20s, is challenged both physically and cognitively. Thus, I was compelled at a young age to study technology and learn how it could improve the lives of everyone—not just the able-bodied.

I chose to study human factors at Loughborough University, and during this time, I gained experience at IBM Hursley and joined the user-centered design team. As a usability engineer intern, I volunteered at HCI conferences to explore potential Ph.D areas of study, ultimately choosing a CS doctoral program at University College London (UCL)—the first university to admit women on equal terms with men. Under the supervision of the esteemed Angela Sasse, I developed a framework for the security industry to optimize operator performance in collaboration with the U.K. Home Office and the Metropolitan Police Service on the design, configuration, and usage of digital CCTV video to detect crime. My time at UCL coincided with the July 7, 2005 bombings and ended up being the reason I was drawn to the field. My research results were timely and hugely impactful to the security industry as well as the public.

Since graduating, I've spent 15 years predominantly within product and design agencies. I chose the industry route over academia as I found university life quite lonely and wanted to work with international businesses. I gained practical experience, working closely with engineers to learn about various applications of technology. I began to understand the complexities, constraints, and dependencies that come with software development.

As a user experience (UX) researcher, I've had the opportunity to conduct thousands of hours of research with customers worldwide, observing user behaviors and workarounds, and translating these findings into rich insights to empower design change. It's been so rewarding to see the impact different products have on peoples' lives. But as organizations started to train teams about product, slowly shifting away from waterfall to agile processes, my level of influence as a researcher became limited.

Over the past five years, I've spent time learning about product management in greater depth. In my last few roles, I transitioned from UX researcher to product manager and now experience design director. As a product lead, my greatest enjoyment has come from positively influencing design. Following formal training and practice, I learned that having a product mindset is essential for organizations as they foster a customer-centric approach, facilitate an iterative product development process, and cultivate a culture of experimentation.

Solving challenges for people and businesses is what I enjoy. I currently head up the insight and data practice at Foolproof, a Zensar product and service design company. Every day is different for me, and I love the variety, change, and challenges in my work. I could be leading research to inform a product strategy, facilitating research to inform ideas and a new product proposition, or leading cross-functional teams in executing customer research, analytics, design, and running experiments to optimize products. I work with a wide range of clients, across many industries, maturity levels, and throughout the product life cycle, which makes my work so interesting. If I look back at my career so far, much of it was intentional due to personal perseverance, passion, and encouragement from my family. The opportunities offered through my professional network didn't hurt either. Technology and people are ever-changing, and the possibilities are endless. By intentionally changing roles, you can forge fulfilling and rewarding career routes. C

acm2y

Supporting Computing Education in Two-Year Higher Education Programs

ACM2Y advocates for a diverse group of computing students by building a targeted and resourceful community for faculty of two-year higher education programs. Its scope encompasses all disciplines in computing education, including computer science, software engineering, information technology, cybersecurity, information systems, data science, and computer engineering.

ACM2Y welcomes all who support computing education at two-year programs. By providing venues for networking, community building, and communication around two-year computing programs, ACM2Y helps two-year college advocates keep abreast of changes in the education landscape that affect two-year programs.

Visit acm2y.acm.org to join the ACM2Y community!

acm Association for Computing Machinery

DOI:10.1145/3596710

The End Is Not Clear

In his January 2023 *Communications* Viewpoint, "The End of Programming," Matt Welsh wrote "nobody actually understands how large AI models work." However, already no one person understands existing large computer systems. Indeed, no team of people understands them. Staff turnover and other practicalities of real life mean not even the team that wrote them originally (should it still exist) nor the team currently responsible for maintaining them, fully understands large software systems, which can now exceed a billion lines of code. And yet such systems are in worldwide daily use and deliver economic benefits.

W.B. Langdon, London, U.K.

Author's response:

It is true that modern software systems are incredibly complex, but fortunately, with the exception of programs written to be intentionally oblique—such as in an esoteric programming language like Malbolge or Whitespace—it is, in fact, possible to directly inspect and understand the source code of such software directly, either by hand or with the aid of automated code analysis tools. A large AI model, on the other hand, is essentially just a bunch of floating point numbers. Just as it is impossible to untangle the operation of the human brain by inspecting the neuronal connections under an electron microscope, AI models are not amenable to direct scrutiny."

Matt Welsh, Seattle, WA, USA

Theory of Relativity

In his response to Robert Glass's comments in the February 2023 *Communications* Letters to the Editor section, Editor-in-Chief James Larus asks the question: Why are more practitioners not members of ACM? I developed software for 44 years and have been a member of the ACM less than half that time. I have mentioned ACM to my workmates and distributed articles I thought would be of interest. My colleagues did not see the relation between the articles and their work. Admittedly, many developers were working on IBM mainframes. Developers had in-depth knowledge of z/OS and its subsystems. I cannot recall any ACM articles that focused on that platform. What does ACM offer those developers that write business applications on z/OS?

I was fortunate to have worked on OS/2, Linux for z, Windows, and z/OS. I have learned, and programmed in, new languages. I have written code that implemented classic algorithms and then found uses for them in my work. I got involved in the Microsoft TEALS program and found the articles on CS in education enlightening. Many practitioners work in the same area for their entire career. Many move into management roles. How would membership in the ACM benefit them?

Personally, I have found membership in the ACM to be of great value. I do not think I am representative of the world of practitioners.

Walter Falby, Detroit, MI, USA

Editor-in-Chief's response:

Stay tuned! The forthcoming Communications *website will allow more space for articles, and* Communications *will use this to publish more material of interest to practitioners.*

James Larus, Editor-in-Chief,
Communications of the ACM
Lausanne, Switzerland

Binary Balancing

Moshe Y. Vardi's May 2023 *Communications* column presents two fundamental—and inseparable—problems. He wants ACM to explicitly commit to the public good.

Problem 1 is there is no agreed-upon definition of the public good. Is pro-abortion part of that good, or is pro-life? Is fare-beating in the subway part of that public good (as the May 2023 *Communications* Law and Technology column by Kendra Albert and James Grimmelmann Viewpoint implies)?

Problem 2 is Vardi attempts to hijack the sole primary purpose of ACM, which is to advance computing.

There are far too many political activists today insisting all groups must support their agenda, or else. This just leads to polarization and divisiveness, which are clearly growing in society. ACM must not let this happen.

Alex Simonelis,
Montreal, Quebec, Canada

Author's response

Alexander Simonelis does not like my claim that the higher purpose of ACM is to serve the public good. He is not the only one. I have been practically accused of being a Marxist-Leninist because of my advocacy of the public good.

In response, I would point out that positioning the public good as the higher purpose of computing as a profession is explicitly stated in the ACM Code of Professional Ethics. Advancing computing as a science and a profession is therefore a means, not an end. Simonelis also bemoans the growing divisiveness of society. But there is strong evidence this divisiveness has, to a significant degree, been brought about by information technology. So Simonelis is essentially saying that to avoid divisiveness we should ignore the role of technology in fomenting societal divisiveness.

I agree with Simonelis when he says we do not have an agreement on how to best serve the public good. As an example, the Preamble of the U.S. Constitution essentially says the higher purpose of the United States is to advance the public good, and the Constitution is the mechanism by which the citizens of the U.S. reach agreement on how to serve the public good. ACM is an association of members, so the members, collectively, will have to decide on how ACM should serve the public good.

Moshe Y. Vardi, Senior Editor
Communications of the ACM
Houston, TX, USA

Communications welcomes your opinion. To contribute a letter to the editor, please limit your comments to 500 words or less and send to letters@cacm.acm.org.

BLOG@CACM

The *Communications* website, https://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:10.1145/3595956 https://cacm.acm.org/blogs/blog-cacm

Unleashing the Power of Deep Learning

The Massachusetts Institute of Technology's Vivienne Sze on how to take greatest advantage of deep learning systems.

Vivienne Sze

Why Businesses Must Untether Deep Learning

March 16, 2023
https://bit.ly/3NwNqYl

While it is clear that deep learning (a core technology used in AI-enabled applications) can deliver tangible results, the technology's applications are still being constrained in several different directions. Many of the limitations in terms of accuracy and ability can be addressed in the coming years as programmers and designers refine their algorithms and pile on more training data.

There is one constraint, though, that seems fundamental to the nature of deep learning, to the extent that many developers almost accept it as the price of doing business: the sheer computation power required to run it.

Tied to the Cloud

Deep learning involves running hundreds of millions or even billions of compute operations (for example, multiplies and additions). The GPT3 model—the foundation for the wildly successful ChatGPT tool—is reported to have used 175 billion parameters, requiring more than 10^{23} compute operations to train (which translates to millions of dollars) and the finished product requires clusters of powerful, and expensive, processers to run effectively.

While this hunger for processing power is not going to surprise anybody familiar with the technology, the limitations it imposes extend far beyond the need to buy more processors. It also makes it extremely difficult to run deep learning on a portable device—the kind of thing that people are likely to have in their home, bag, or pocket.

The GPT3 model is reported to have used 175 billion parameters, requiring more than 10^{23} compute operations to train (which translates to millions of dollars).

This need for processing does not mean we cannot access AI tools on our mobile devices, of course. You can easily open a Web browser on your iPad and use it to produce all the AI-generated art, email, and stories you could ever want, but the resource-heavy computation is handled in the cloud rather than locally. This cloud-based setup means any device, tool, or application looking to take advantage of deep learning must be tied to the Internet. Unfortunately, it is not always feasible to have a reliable connection. Depending on where you are in the world, you may have limited bandwidth. Even if high-bandwidth 5G connectivity is an option, you may still face the kind of reliability issues familiar to anyone who has tried to use AI-enabled speech recognition to switch songs on a cross-country road trip.

There are, of course, plenty of applications where a permanent connection to the cloud is not too much of an ask. For example, the AI used for automation at a factory is unlikely to be

PHOTO BY LILLIE PAQUETTE

installed somewhere without a decent Internet connection. However, there are a huge number of use cases where we could benefit from local AI (such as performing the computation locally, where data is collected, rather than in the cloud). For those working with sensitive data—for instance, health or financial data—sending the data to the cloud may also pose security or privacy concerns. For those working on applications that require a fast reaction time—for instance, autonomous navigation or augmented/virtual reality—the time it takes to send the data to cloud and wait for a response may impose too large of a latency. To achieve local AI, we need to change how we think about designing both our processing hardware and our deep learning software.

The Limits of Thought

Before we make the mistake of assuming the solution is to increase the amount of processing hardware to throw at the problem, it is important to remember that a lack of raw processing speed is not the only factor that makes portable deep learning challenging. Even if we were somehow able to miniaturize one of the powerful racks used in cloud computing set-ups, we would still have to contend with two associated issues: heat, and energy consumption.

These are both perennial challenges for any computing set-up, but they are amplified for smaller, portable devices. A self-driving car can use more than 1,000 watts to compute and analyze the data from its cameras and other sensors, for example. In contrast, a typical smartphone processor must run at approximately one watt. The more power the processor consumes, the more heat it produces. Processors in the cloud can generate so much heat that they require liquid cooling. This would not be feasible for a small device trying to handle deep learning applications in your pocket or your backpack; not only is liquid cooling infeasible, but often you can't even use a fan, due to the size and weight constraints of these portable devices.

Even if we could get sufficient power into the processor, we still must answer the question of energy storage. What's the point in having an AI assistant embedded in your smartphone if you can only run it for 10 minutes before having to find an outlet to recharge your battery?

For those working on applications that require a fast reaction time—for instance, autonomous navigation or augmented/virtual reality—the time it takes to send the data to the cloud and wait for a response may impose too large of a latency.

Untethering AI

If developers and designers want to embrace all the benefits of untethered AI—whether that means a semi-intelligent delivery drone or a camera capable of real-time image processing—we need to eliminate these challenges. Or, at least, find a way to mitigate their impact.

The answer comes in the form of improved efficiency at every possible level of the equation, from the code we use to specify the deep learning operations through to the hardware we run it on. It's possible to write algorithms that take our existing deep learning concepts but run them more efficiently, both in terms of computation and power draw. For example, we are able to trim areas of code that were valuable during the AI learning and training process but serve little use in a finished product.

On the hardware side of things, there are plenty of ways to make AI and deep learning technology run more efficiently. Most of the processor hardware we currently use in our computers to run our software are generalist—they are pretty good at handling any application you throw at them but do not excel at any individual task. Yet it is possible to design hardware specially tailored to the demands of deep learning. Google released a chip specialized for the matrix operations in deep learning more than half a decade ago (https://bit.ly/3Hw5Nca), while my own team at MIT has developed one built specifically for deep learning tasks (https://bit.ly/3NAm1F0) that can operate on less than one-third of a watt and is 10 times more energy-efficient than a more general purpose mobile processor.

We find the greatest opportunities for improvement when we begin to think about hardware and software at the same time. When these two aspects work in tandem, we can apply techniques to eliminate unnecessary computations and reduce the distance that information needs to travel across the physical chip during computations; both approaches can speed up the chip and reduce its power and energy consumption.

If carefully applied, these advancements can untether the power of deep learning, and allow businesses to truly innovate with intelligent devices.

A Wide-Ranging Benefit

The focus of my work is on being able to overcome the hurdles that keep AI technology tethered to the cloud, but the benefits of making deep learning more efficient do not stop there.

Datacenters, for example, are already infamous for their heavy power draw. As AI usage becomes more and more mainstream, it is not hard to predict that a significant chunk of this power will be devoted to running deep learning algorithms. If we can develop the technology to run these systems more efficiently, the energy savings have the potential to be massive, including reducing its carbon footprint and helping us move towards a more sustainable future.

These energy savings upstream can also mean cost savings downstream, making it easier for more users—with even more modest budgets—to access the incredible power of deep learning. With the age of AI descending upon us, this can only be a good thing.

Vivienne Sze is associate professor of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology (MIT), co-author of the book *Efficient Processing of Deep Neural Networks*, and lead instructor of the MIT Professional Education course *Designing Efficient Deep Learning Systems*.

Science | DOI:10.1145/3595957 Chris Edwards

Accelerating Optical Communications with AI

Communications efficiency helps bring photonics to artificial intelligence.

PHOTONIC COMPUTING HAS seen its share of research breakthroughs and deep research winters, much like the history of artificial intelligence (AI). Now, with the resurgence of AI, the huge amounts of energy today's large neural-network models need when running on electronic computers is reawakening interest in uniting the two.

More than 30 years ago, during one of the booms in research into artificial neural networks, Demetri Psaltis and colleagues at the California Institute of Technology demonstrated how techniques from holography could perform rudimentary face recognition. The team members showed they could store as many as one billion weights for a two-layer neural network using the core elements from a liquid-crystal display. Similar spatial light modulators became the foundation of several attempts to commercialize optical computing technology, including those by U.K.-based startup Optalysys, which has focused in recent years on applying the technology to accelerating homomorphic encryption to support secure remote computing.

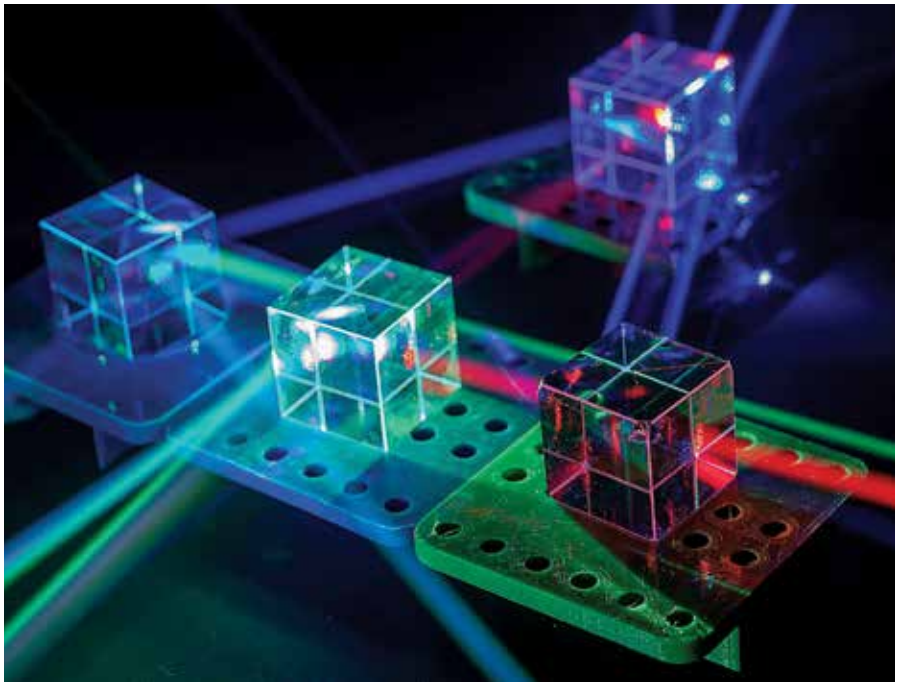


IMAGE FROM SHUTTERSTOCK.COM

Though some groups are using spatial light modulators for AI, they represent just one category of optical computer suitable for the job. There are also decisions as to which form of neural networks best suits optical computing. Some techniques focus on the matrix-arithmetic operations of mainstream deep learning pipelines, while others focus squarely on emulating the spike trains of biological brains.

What all the proposed systems have in common is the possibility that, by using photons to communicate and calculate, they will deliver major advantages in density and speed over systems based purely on

electrical signaling. A 2021 study of inferencing based on matrix arithmetic by Mitchell Nahmias, now CTO of startup Luminous Computing, and colleagues at Princeton University argued the theoretical efficiency of AI inferencing in the optical domain could far surpass that of conventional accelerators based on existing electronics-only architectures.

The key issue for machine learning is the amount of energy needed to move data around accelerators. Electronic accelerators often employ strategies to cache as much data as possible to reduce this overhead, with major trade-offs concerning whether temporary results or weights are held in the cache depending on the model's structure. However, the energy cost of delivering photons over distances larger than the span of a single chip is far lower than it is for electrical signaling.

A second potential advantage of photonic AI comes from the ease with which it can handle complex operations in the analog domain, though the power savings achievable here are less certain than for communications. Whereas matrix arithmetic relies on highly parallelized hardware circuits for performance in conventional systems, simply passing photons through an optical component such as a Mach-Zehnder interferometer (MZi) or micro-ring resonator will perform arithmetic requiring hundreds or even thousands of logic gates in a digital circuit.

In the MZi, coherent beams of light pass through a succession of couplers and phase shifters. At each coupling point, the interference between the beams results in phase shifts that can be interpreted as part of a matrix multiplication. A 4x4 matrix operation requires just four inputs that feed into six coupling elements, with four output ports providing the result. The speed of the operations is limited only by the rate at which coherent pulses can be passed through the array.

In analog architectures, noise presents a significant hurdle. Work by numerous groups on accelerating inference operations has shown that deep neural networks can work successfully at an effective resolution of 4 bits, but hardware overhead and energy rise quickly as resolution increases. These effects may limit the practical energy advantage of photonic designs.

Estimates by Alexander Tait, assistant professor at Queen's University, Ontario, found the potential power-savings easily eroded by the practical limitations of today's optical devices. Tait calculated just 500 micro-ring resonators acting as neurons in a fully connected layout could fit onto a single $1cm^2$ die using early 2020s technology. But operating at 10GHz, it would require a kilowatt of power. Tait stresses the example shows the impact of the current need for heaters to tune optical properties. Scaling and design changes could bring the energy down dramatically. "The heaters are certainly a solvable problem," he says.

One possible direction is to use phase-change materials like those used in rewritable DVDs and non-volatile memories. These materials exhibit changes in their electro-optical properties based on how quickly they are cooled after rapid heating. But as they do not need constant heating, they would result in lower-energy optical computers if they can be made to work reliably.

"If you have a phase shifter that is low energy, there are so many things you can do in integrated photonics," says Carlos Ríos Ocampo, professor of materials science and engineering at the University of Maryland.

However, there are differences in how the materials can be applied. Tait points out that while the energy of MZi and micro-ring multiplier implementations are similar today, because the tuning needed for MZi devices is dynamic, the non-volatile nature of phase-change materials does not convey as big a benefit as for micro-rings.

"If you have a phase shifter that is low energy, there are so many things you can do in integrated photonics."

Another factor in favor of optical processing is that the high bandwidth and speed of photonic processing means integration need be nowhere the level of electronic processing.

"The metric is compute density. If you do vector or matrix operations with photonics you don't need dense core devices: with the higher speed you can use the hardware recursively," says Bhavin Shastri, assistant professor at Queen's University.

A potentially more troublesome problem for photonic neural networks lies in the need for non-linear behavior in AI algorithms. Most of the conventional photonic components work entirely linearly and lack the flexibility of transistors that can operate in linear and non-linear regions.

"Non-linear interactions become quite challenging. Either you need new materials to enhance the non-linearities or go to optical-electronic-optical conversion, and this has to be done in a very efficient way," Shastri says. Today, the delay in electro-optical conversion reduces the theoretical throughput advantage of photonic AI.

Ríos Ocampo points out that the semiconductor fab operators who could build the necessary silicon photonics chips are reluctant to introduce novel materials into their processes without a large market driver. Photonic computing has yet to provide the necessary demand. Some help may come from the experience semiconductor companies have obtained in their decades-long work on electronic phase-change memories. This could assist the development of optical non-volatile memories and other components that will be useful in these systems, though most of the focus in semiconductors has been on compounds that absorb light. Transparent materials that can be used to manipulate phase alone would need far more testing for compatibility with semiconductor fab equipment.

Despite photonic AI being at an early stage of its evolution, several start-ups have embarked on plans to build commercial photonic accelerators,

taking advantage of improvements in silicon-photonics made so far, primarily in response to the needs of the networking and communications sector. Communication throughput remains a key focus in the AI world as well. Among the small group of startups, Lightmatter is working on both a photonic accelerator and an optical interconnect technology that can be used to improve communications speed between electronic modules.

Luminous Computing originally planned to build its own photonic AI system, but the company has opted to focus on an electronic accelerator supported by a photonic interconnect of its own design.

Luminous president Michael Hochberg says having investigated its options for a photonic core, "We concluded that the bottlenecks that were preventing dramatic improvements were elsewhere in the system."

A brighter future for full photonic AI may lie outside datacenter systems, such as those being built by Lightmatter and Luminous. "This other community is looking at applications where electronics fundamentally has challenges," Shastri says. "There are some things you can't just offload to a data center. The question then becomes: What are those tasks?"

Though the trend was short-lived at the time, the driver behind research into optical-only routers in the late 1990s was the belief that it would be easier to steer packets at line speed by keeping the data in the photonic domain and not take the hit of electro-optical and digital-analog conversion. There are numerous possible applications where sensor outputs can be taken and used directly without having to go through analog-digital conversion, in contrast to datacenter systems that work on stored data.

Shastri points to recent work on using analog photonic processing to separate radio transmissions. "It's a form of the cocktail-party problem: In crowded radio spectrum you need to find and focus on just one signal and use intelligent ways of processing it. You can't just use filtering. The bandwidth of electronics is narrow and the scaling of energy if you use electronic processing can grow quadratically depending on the number of channels or antennas. We showed you can process at really wide bandwidths. And the energy scales linearly."

"There are some things you can't just offload to a datacenter. The question then becomes: What are those tasks?"

Another potential application lies in optimization tasks as well as predictive control and stabilization for fast-moving vehicles, such as hypersonic aircraft. "Where you need to converge to a solution really fast, photonics can have an advantage."

Though much hinges on the ability to integrate multiple technologies in a cost-effective way, the massive demand for data intelligence in a growing range of applications may provide photonic computing with the conditions needed to avoid another R&D winter. C

Further Reading

Huang, C. et al.
Prospects and Applications of Photonic Neural Networks
Advances in Physics: X, 7:1, 1981155 (2021)

Nahmias, M.A, Ferreira de Lima, T., Tait, A.N., Peng, H.-T., Shastri, B.J, and Prucnal, P.R.
Photonic Multiply-Accumulate Operations for Neural Networks
IEEE Journal of Selected Topics in Quantum Electronics, Volume 26, Issue 1 (2020)

Ríos Ocampo, C.A. et al.
Ultra-Compact Nonvolatile Photonics Based on Electrically Reprogrammable Transparent Phase-Change Materials
PhotoniX, 3:26 (2022)

Tait, A.N.
Quantifying Power in Silicon Photonic Neural Networks
Physical Review Applied 17, 054029 (2022)

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

ACM Member News

AT THE INTERSECTION OF HCI, ML, AND UBIQUITOUS COMPUTING

"My brother had a Timex Sinclair computer, and I got really excited about it," recalls Anind Dey, professor and dean of the Information School at the University of Washington in Seattle, WA.

This early interest led him to earn his undergraduate degree in Computer Engineering from Simon Fraser University in Burnaby, Canada. Dey later received master's degrees in aerospace engineering and computer science from the Georgia Institute of Technology (Georgia Tech) in Atlanta, GA, where he also obtained a Ph.D. in computer science in 2000.

Dey then went to work at the Intel Research Lab at the University of California, Berkeley. He remained there until 2004, when he joined the faculty of the Human-Computer Interaction Institute at Carnegie Mellon University in Pittsburgh, PA. Dey moved to the University of Washington in 2018, where he has remained ever since.

Today, Dey's research focuses on human-computer interaction, machine learning, and ubiquitous computing.

"My early work centered on context-aware computing, where computational processes are aware of the context in which they operate and can adapt appropriately to that context," he recalls.

For the past few years, Dey has focused on passively collecting large amounts of data about how people interact with their phones and the objects around them, to use in producing detection and classification models for human behaviors of interest.

"I'm now focused on effectively improving the use of information—the tools, algorithms, and interaction with people—to solve problems," Dey says.

—John Delaney is a freelance technology writer based in Manhattan, NY, USA.

Technology | DOI:10.1145/3596206 Neil Savage

The Rise of the Chatbots

How do we keep track of the truth when bots are becoming increasingly skilled liars?

During the 2016 U.S. presidential race, a Russian "troll-farm" calling itself the Internet Research Agency sought to harm Hillary Clinton's election chances and help Donald Trump reach the White House by using Twitter to spread false news stories and other disinformation, according to a 2020 report from the Senate Intelligence Committee. Most of that content apparently was produced by human beings, a supposition supported by the fact that activity dropped off on Russian holidays.

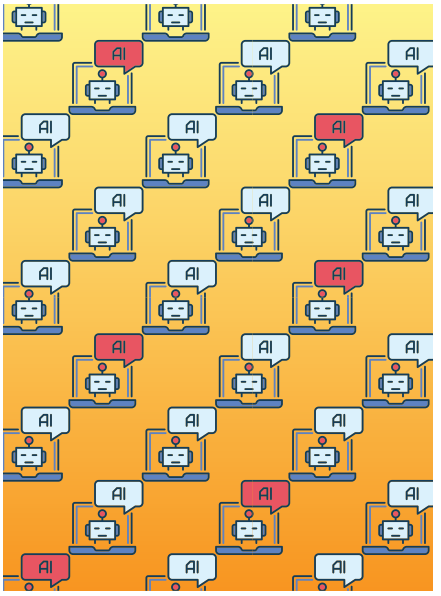
Soon, though, if not already, such propaganda will be produced automatically by artificial intelligence (AI) systems such as ChatGPT, a chatbot capable of creating human-sounding text.

"Imagine a scenario where you have ChatGPT generating these tweets. The number of fake accounts you could manage for the same price would be much larger," says V.S. Subrahmanian, a professor of computer science at Northwestern University, whose research focuses on the intersection of AI and security problems. "It'll potentially scale up the generation of fakes."

Subrahmanian co-authored a Brookings Institution report released in January that warned the spread of deepfakes—computer-generated content that purports to come from humans—could increase the risk of international conflict, and that the technology is on the brink of being used much more widely. That report focuses on fake video, audio, and images, but text could be a problem as well, he says.

Text generation may not have caused problems so far. "I have not seen any evidence yet that malicious actors have used it in any substantive way," Subrahmanian says. "But every time a new technology emerges, it is only a matter of time, so we should be prepared for it sooner rather than later."

There is evidence that cybercriminals are exploring the potential of text generators. A January blog post from security software maker Checkpoint said that in December, shortly after ChatGPT was released, unsophisticated programmers were using it to generate software code that could create ransomware and other malware. "Although the tools that we present in this report are pretty basic, it's only a matter of time until more sophisticated threat actors enhance the way they use AI-based tools for bad," the company wrote.



Meanwhile, Withsecure, a Finnish provider of cybersecurity tools, warned of the threat of so-called "prompt engineering," in which users coax software like ChatGPT to create phishing attacks, harassment, and fake news.

ChatGPT, a chatbot based on a large language model (LLM) developed by AI company OpenAI, has generated much excitement as well as fear about the advances of AI in general, and there has been a backlash from many technologists across varying disciplines. There have been calls to pause AI's development, and at press time, a one-sentence open letter to the public signed by hundreds of the world's leading AI scientists, researchers, and more (including OpenAI CEO Sam Altman) warned that "[m]itigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war. Microsoft, which invested in its development, soon incorporated the chatbot into its search engine, Bing, leading to reports of inaccurate and sometimes creepy conversations. Google also put out a version of Bard, its own chatbot based on its LaMDA LLM that had previously made the news when a Google engineer proclaimed it was self-aware (he was subsequently fired).

Despite some early misfires, the text generated by these LLMs can sound remarkably like it was written by humans. "The ability to generate wonderful prose is a big and impressive scientific accomplishment from the ChatGPT team," Subrahmanian says.

Fake Detectors

Given that, researchers agree it would be useful to have a way to distinguish human-written text from that generated by a computer. Several groups have developed detectors to identify synthetic text. At the end of January, OpenAI released a classifier designed to distinguish between human and machine authors, hoping to both identify possible misinformation campaigns and cut down on the risk of students using a text generator to cheat on their schoolwork. The company warns its classifier is not fully reliable; in tests, it labeled 9% of human-written texts as AI-written, was unreliable on texts of less than 1,000 characters, and did not work well in languages other than English.

Bimal Viswanath, a professor of computer science at Virginia Tech, says some detectors demonstrated very high accuracy when their developers tested them on synthetic text that those developers had generated, but did less well with fake text found in the real world, where the distribution of data may be different from what was created in the laboratory and where malicious actors try to adapt to defenses.

IMAGE BY VENOMOUS VECTOR

AI-written text is thought to be detectable because of the way it is created. The LLMs are trained on human-written text and learn statistics about how often particular words appear in proximity to other words. They then make predictions about how likely it is that a given word is the best choice to appear next in a sentence and pick the word with the highest probability, generally speaking. Humans show more diversity in their choices of words, and that difference in diversity can be perceived.

Viswanath underscores the difficulty of being able to say for sure why detectors peg a particular text as real or fake. They use neural networks and deep learning to identify hidden patterns in sequences of text, but as with so much of deep learning, scientists cannot always identify the patterns. Attackers also can evade the detectors by modifying their language generator; having it select slightly fewer high-probability words, for instance, can introduce enough randomness in the word choice to make the text seem human-generated to a neural network.

That strategy has its limitations. If a malicious actor is trying to get out a particular message, they cannot change the text so much that that message is lost. "You have a certain thing that you want to communicate. You don't want to change that underlying semantic content," Viswanath says. That points to a method that might be better at detecting fake text. Because the LLM does not really know what it is talking about, it can inadvertently select words with different meanings. For instance, it might start talking about named places or people, but within a few sentences it could drift into a different set of names. "And then the article may not sound coherent anymore," he says. Using semantic knowledge to detect synthetic text, though, is an area that still requires a lot of research, he adds.

Watermarking

Another approach to identifying synthetic text is building in a hidden pattern when the text is created, a process known as watermarking. Tom Goldstein, a professor of computer science at the University of Maryland, has developed a scheme to embed such a pattern in AI-generated text. His system uses a pseudo-random number generator to assign each token in a text—a character or sequence of characters, often a single word—to either a red list or a green list. Humans, not knowing which list a word is on, should choose a roughly equal proportion of red- and green-list words within a mathematically predictable variation.

The text generator, meanwhile, assigns extra weight to green-list words, making them more likely to be chosen. A detector that knows the algorithm used to generate the list, or even just the list itself, then examines the text. If it's near half-red and half-green, it decides a human wrote it; if the green words greatly outweigh the red, the machine gets the credit.

It only takes 36 tokens—approximately 25 words—to produce a very strong watermark, Goldstein says, so even individual tweets can be labeled. On the other hand, it is possible to weaken or remove a watermark by having a human or another LLM rewrite the text to include more red-list words. "The question is, how much of a sacrifice in quality do you need to suffer to remove the watermark?" Goldstein says.

In fact, Viswanath says, every defense can be defeated, but at a cost. "If you've raised the cost of the attack so significantly that the attack is no longer worth it, then you've actually won as a defender," he says.

Aside from deliberate misuse, text generators also can generate toxic content unintentionally. Soroush Vosoughi, a professor of computer science in the Institute for Security, Technology and Society at Dartmouth University, is working on methods for countering the antisocial possibilities of text generation by looking for ways to make chatbots prosocial. "We develop models that can sit on top of these language models and guide their generation," he says.

For instance, Vosoughi has developed a classifier, based on ratings from groups such as the Pew Research Center that classify news outlets as leaning left or right politically. The classifier learns to identify certain words as more indicative of a political bias and steers the chatbot to give greater weight to neutral terms. It might, for instance, push the generator away from following the word "illegal" with "aliens" and instead encourage it to write "immigrants." Another version waits until the whole sentence is generated, and then can go back and change the phrase to "undocumented immigrants." The same sort of approach can be used with, say, medical information, to make it less likely the generator will produce misleading advice.

Of course, this approach requires humans to define the values they want the LLM to uphold, Vosoughi says, but at least it can avoid the problem of models inadvertently generating hate speech or misinformation.

None of these solutions is permanent, researchers warn. Every success at labeling or detecting machine-written text is likely to be met by more sophisticated methods of evading such detection. That does not mean sitting out such an arms race is an option, Vosoughi says. "We need to be just one step ahead of the other side," he says. "That's the best we can do in these situations." C

Further Reading

Pu, J., Sawar, Z., Abdullah, S. M., Rehman, A., Kim, Y., Bhattacharya, P., Javed, M., and Viswanath, B.
Deepfake Text Detection: Limitations and Opportunities, IEEE Symposium on Security and Privacy 2023.
https://doi.org/10.48550/arXiv.2210.09421

Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T.
A Watermark for Large Language Models, 2023, arXiv, https://doi.org/10.48550/arXiv.2301.10226

Liu, R., Jia, C., Wei, J., Xu, G., Wang, L., and Vosoughi, S.
Mitigating Political Bias in Language Models Through Reinforced Calibration, 2021, *Proc. of the AAAI*, https://doi.org/10.48550/arXiv.2104.14795

Byman, D.L., Gao, C., Meserole, C., and Subrahmanian, V.S.
Deepfakes and International Conflict, 2023, Foreign Policy at Brookings, https://www.brookings.edu/research/deepfakes-and-international-conflict/

What is ChatGPT? OpenAI's ChatGPT Explained https://www.youtube.com/watch?v=o5MutYFWsM8

Neil Savage is a science and technology writer based in Lowell, MA, USA.

Society | DOI:10.1145/3595958 Gregory Mone

Outsmarting Deepfake Video

Looking for truth, when "it's becoming surprisingly easy to distort reality."

In March 2022, a synthesized video of Ukrainian President Volodymyr Zelenskyy appeared on various social media platforms and a national news website. In the video, Zelenskyy urges his people to surrender in their fight against Russia; however, the speaker is not Zelenskyy at all. The minute-long clip was a deepfake, a synthesized video produced via deep learning models, and the president soon posted a legitimate message reaffirming his nation's commitment to defending its land and people.

The Ukrainian government already been had warning the public that state-sponsored deepfakes could be used as part of Russia's information warfare. The video itself was not particularly realistic or convincing, but the quality of deepfakes has been improving rapidly. "You have to be a little impressed by synthetic media," says University of California, Berkeley computer scientist and digital forensics expert Hany Farid. "In five years, we have gone from pretty cruddy, low-resolution videos to full-blown, high-resolution, very sophisticated 'Tom Cruise TikTok' deepfakes. It's evolving at light speed. We're entering a stage where it's becoming surprisingly easy to distort reality."

In some cases, such as the aforementioned TikTok example, in which a company generated a set of videos that closely resemble the famous actor, the result can be entertaining. Startups are engineering deepfake technology for companies to use in marketing videos, and Hollywood studios are slipping hyper-realistic digital characters into movies alongside human actors. Yet the malicious use of this technology for disinformation, blackmail, and other unsavory ends is concerning, according to researchers. If the deepfake of Zelenskyy had been as realistic as one of those Tom Cruise clips, the synthesized video could have had terrible consequences.

A deepfake video of what appears to be Ukraine president Volodymyr Zelenskyy telling his countrymen to lay down their weapons and surrender to Russia.

The potential for nefarious applications, and the pace at which the deepfake techniques are evolving, has sparked a race between the groups generating synthetic media and the scientists working to find more effective and resilient ways of detecting them. "We're playing this chess game in which detection is trying to keep pace or advance ahead of creation," says computer scientist Siwei Lyu of the University at Buffalo, State University of New York. "Once they know the tricks we use to detect them, they can fix their models to make the detection algorithms less effective. So every time they fix one, we have to develop a better one."

The roots of deepfake technology can be traced back to the development of generative adversarial networks (GANs) in 2014. The GAN approach pits two models against each other. In their paper introducing the concept, Ian Goodfellow and colleagues described the two models as analogous to the "game" between counterfeiters and police; the former tries to outwit the latter, and competition pushes them to a point at which the counterfeits approach the real thing. With deepfakes, the first model generates a synthetic image and the second tries to detect it as a fake. As the pair iterate, the generative model corrects its flaws, resulting in better and better images.

In the early days, it was relatively easy for people to recognize a fake video; inconsistencies in skin tone or irregularities in facial structure and movement were common. As the synthesis engines have improved, however, detection has become increasingly difficult. "People often think they're better than they are at detecting fake content. We're falling for stuff, but we don't know it," says Sophie Nightingale, a psychologist studying deepfake recognition at Lancaster University in the U.K. "We're arguably at the point where the human perceptual system can't tell if something is real or fake."

To keep pace with the evolution of the technology, researchers have been developing tools to spot telltale signs

SCREEN CAPTURE FROM VIDEO. COURTESY OF THE TELEGRAPH/YOUTU.BE/X17YREV5SL4

of digital forgery. In 2018, ACM Distinguished Member Lyu and one of his students at the University at Buffalo were studying deepfake videos in hopes of building better detection models. After watching countless examples and using publicly available technology to generate videos of their own, they noticed something strange. "The faces did not blink!" Lyu recalls. "They did not have realistic blinking eyes, and in some cases they did not blink at all."

Eventually, they realized the lack of eye blinking in the videos was the logical outcome of the training data. The models that generate synthetic video are trained on still images of a given subject. Typically, photographers do not publish images in which their subjects' eyes are closed. "We only upload images with open eyes," Lyu explains, "and that bias is learned and reproduced."

Lyu and his student created a model that detected deepfakes based on the lack of eyeblinks or irregular patterns in eye blinking, but not long after they released their results, the next wave of synthetic videos evolved. The Zelenskyy video, while poor in quality, does feature the Ukrainian president blinking.

The eye-blinking work is reflective of the predominant approach to detecting deepfakes: searching for evidence or artifacts of the generative or synthetic process. "These generative models learn about the subjects they recreate from training data," says Lyu. "You give them lots of data and they can create realistic synthetic media, but this is an inefficient way of learning about the real world, because anything that happens in the real world has to follow the laws of the real physical world, and that information is indirectly incorporated into the training data." Similarly, Lyu has pinpointed inconsistencies between the reflections in the corneas of the eyes of synthesized subjects and nearly imperceptible differences in the retinas.

The deep learning researcher Yuval Nirkin, currently a research scientist at CommonGround-AI, developed a detection method that compares the internal part of the face in a video with the surrounding context, including the head, neck, and hair regions. "The known video deepfake methods don't change the entire head," sayss Nirkin. "They

Eventually, they realized the lack of eye blinking in the videos was the logical outcome of the training data.

focus only on the internal part of the face because while the human face has a simple geometry that is easy to model, the entire head is very irregular and contains a lot of very fine details that are difficult to reconstruct." Nirkin developed a model that segments a subject's face into inner and outer portions and extracts an identity signal from each. "If we find a discrepancy between the signals of the two parts," he explains, "then we can say someone altered the identity of the subject." The advantage of this approach, Nirkin adds, is that it is not focused on the flaws or artifacts associated with one particular deepfake generation model and can thereby be applied to unseen techniques.

At the University of California, Berkeley, Farid is pioneering a detection method that moves even further away from the focus on specific artifacts. Instead of looking for unrealistic signals, Farid and his students flipped the task around and designed a tool that studies actual, verified video footage of a person. The group's solution hunts for correlations between 780 different facial, gestural, and vocal features within that footage to build a better model of a particular person and that subject's facial, speech, and gestural patterns. Turning your head while you are speaking, for example, will change your vocal tract and generate slight changes in the sound of your voice, and the model identifies such links. As for Zelinskyy, among other things, he has a specific kind of asymmetry to his smile and certain habits of moving his arms as he speaks.

The researchers aggregate all these observations and correlations to create a model or classifier of the famous person, such as Zelinskyy. The accuracy of the classifier increases as more correlations are incorporated, reaching a 100% success rate when the group factored in all 780. When the classifier studies a video, and multiple features fall outside the model, then the technology concludes the sample is not actually the subject. "In some ways, we're not building a deepfake detector," Farid explains. "We're building a Zelinskyy detector."

Farid recognizes the synthesis engines are constantly improving; his group is not publicly releasing the code behind its classifier in hopes of slowing that evolution. Currently they are expanding their database and creating detectors for more world leaders.

As the deepfake generators improve, and the line between real and synthetic media becomes increasingly difficult to discern, developing new means of quickly detecting them only becomes more important. "Getting that balance right and making sure people rely on and trust the things they should and distrust the things they should not, that is a difficult but critical thing to do," explains Nightingale, the psychologist and deepfake researcher. "Otherwise, we could end up in a scenario where we do not trust anything." C

Further Reading

Goodfellow, I. et al.
"Generative adversarial networks," *Communications*, Volume 63, Issue 11, November 2020.

Nightingale, S. and Farid, H.
"AI-synthesized faces are indistinguishable from real faces and more trustworthy," *PNAS*, February 14, 2022.

Boháček, M. and Farid, H.
"Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms," *PNAS*, November 23, 2022.

Nirkin, Y. et al.
"Deepfake Detection Based on Discrepancies Between Faces and Their Context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, October 2022.

Li, Y., Chang, M., and Lyu, S.
"In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," *IEEE Workshop on Information Forensics and Security* (WIFS), Hong Kong, 2018.

Gregory Mone is the co-author, with Daniela Rus, of the forthcoming book *The Heart and the Chip.*

DOI:10.1145/3597151

Pamela Samuelson

Legally Speaking
Legal Challenges to Generative AI, Part I

Questioning the legality of using in-copyright works for training data and producing outputs derived from copyrighted training data.

Generative artificial intelligence (AI) has captured considerable popular attention recently. ChatGPT and DALL-E have given members of the general public opportunities to use AI systems to generate text and image outputs for fun and a wide range of other purposes. Google and Meta have announced their intentions to launch similar AI systems soon.

Generative AI has also caught the attention of lawyers who question the legality of ingesting in-copyright works as training data and producing outputs derived from copyrighted training data.

Lawyers representing four programmers (identified so far as John Does) have, for instance, sued GitHub, Microsoft, and OpenAI, alleging that GitHub's Copilot and OpenAI's Codex AI programs have violated laws by using publicly available open source code (including programs the Does developed) posted on GitHub's site as training data for their generative AI systems. Also illegal, say these Does, are Copilot and Codex outputs of code sequences in response to user prompts insofar as the sequences are substantially similar or virtually identical to open source code used as training data.

The Does claim to represent a class of programmers whose legal rights GitHub and OpenAI have violated. They want a federal court to issue an injunction against these generative AI systems and to award the class $9 billion in statutory damages.

OpenAI developed Codex as a generative AI model trained on billions of lines of publicly available computer source code.

This column focuses on the *Doe v. GitHub* lawsuit as the first of a two-part series on legal challenges to generative AI. A subsequent column will address two similar lawsuits brought against Stability AI for its use of images as training data and for producing outputs based on the training data.

Background on Codex and Copilot

GitHub is an Internet hosting service for software development and version control. It reports having more than 100 million registered developers and hosting 372 million code repositories, including 28 million public repositories. Microsoft acquired GitHub for $7.5 billion in 2018.

OpenAI developed Codex as a generative AI model trained on billions of lines of publicly available computer source code, including code available in GitHub's public repositories. Codex discerns statistical patterns in the structure of existing code. It infers these patterns based on a complex probabilistic analysis of the training data. In response to a user's prompt, Codex produces code to implement the desired function.

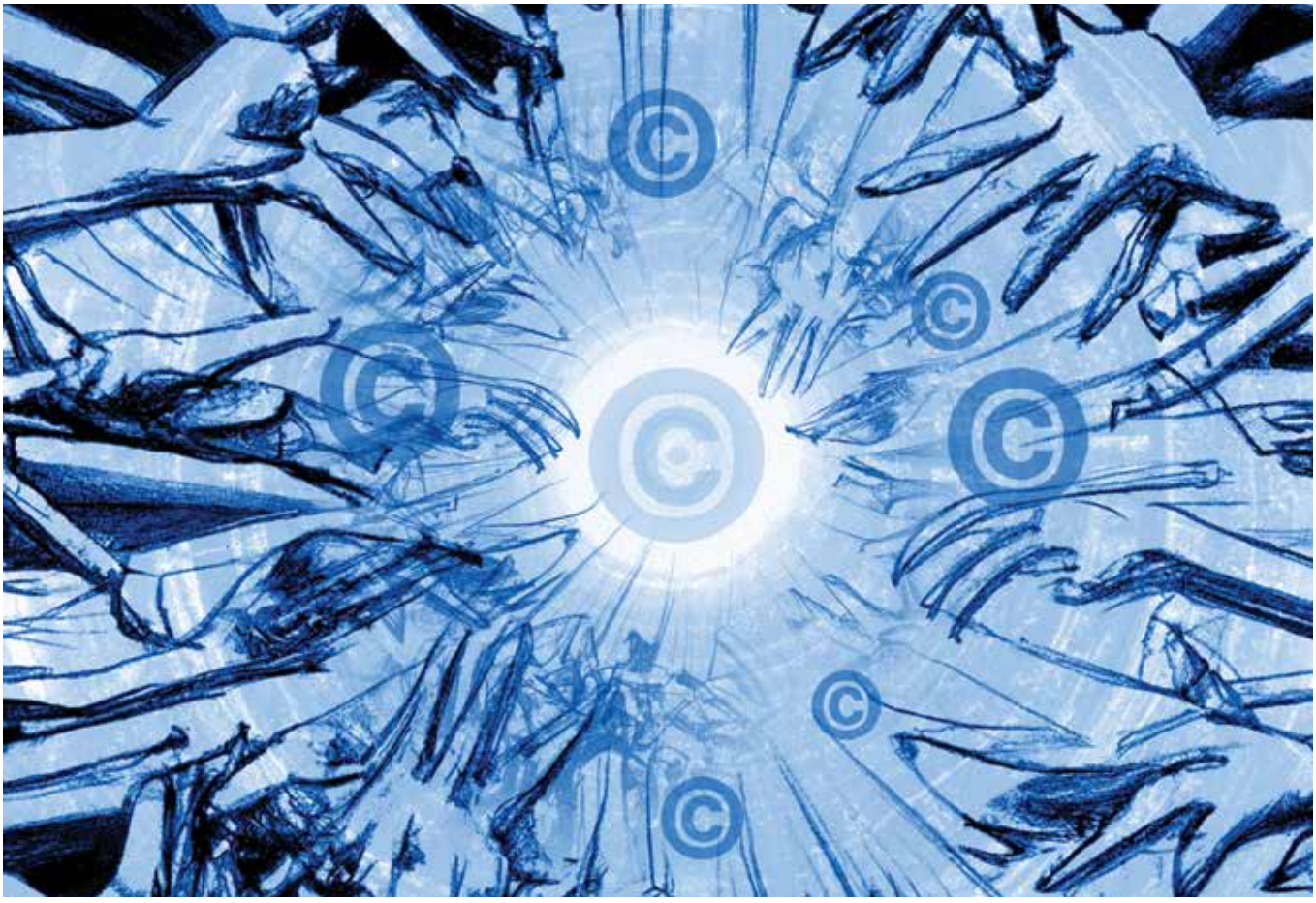


IMAGE CREATED BY ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

In June 2021, GitHub and OpenAI launched Copilot as a cloud-based AI technology that uses Codex to assist the development of software. GitHub users can install Copilot as an extension to various code editors. Copilot treats a user's input to a code editor as a prompt and generates suggested code that may be suitable for the developer's purposes. Copilot subscriptions are available to GitHub users for $10 per month or $100 per year.

What Laws Were Arguably Violated?

Although the complaint says Copilot and Codex are engaged in "software piracy on an unprecedented scale," it does not actually claim GitHub or OpenAI have infringed any copyrights. This is curious because open source software is generally protected by copyright and copyright is the legal basis on which open source licenses are predicated.

The Does' most significant claim is that Copilot and Codex wrongfully removed copyright notices and other copyright-relevant information from open source programs ingested as training data.

The intentional removal or alteration of copyright management information (CMI) from copies of copyrighted works with knowledge that the removal or alteration of CMI is likely to induce, enable, facilitate, or conceal copyright infringements is illegal under § 1202 of Title 17 of the U.S. Code.

A second principal claim is that GitHub and OpenAI have breached open source license agreements by failing to respect license terms, such as requirements to give attribution to the open source developers whose code has been ingested and is being used to generate outputs and to include copyright notices in reused code.

The Does charge GitHub and OpenAI with several other legal violations, including misrepresenting others' licensed code as their own, fraud, unjust enrichment, and violating California unfair competition and privacy laws. This column omits discussion of these subsidiary claims because the lawsuit will primarily be focused on the two principal claims.

Motions to Dismiss

Lawyers initiate lawsuits by filing complaints explaining the legal theories on which the lawsuits are based and core facts that support those legal theories. If courts uphold at least one theory in a case, the plaintiffs may be eligible for certain remedies, such as injunctions and damages.

After reviewing complaints, defendants' lawyers sometimes decide to file motions to dismiss complaints for failure to state claims on which courts could grant the remedies requested in the complaint.

When considering motions to dismiss complaints, courts assume that all of the facts stated in the complaint are true (even if the defendants' lawyers plan to contest their truthfulness if the court denies their motions).

Instead of filing an answer to the Does' complaints, which would typically admit some allegations, deny others, and raise defenses, GitHub and OpenAI filed motions to dismiss the Does' complaints for failure to state claims on which relief could be granted.

Among other things, GitHub and

OpenAI point out the Does have not identified any code in which they claim rights. Nor have they specified any injury they suffered as a result of GitHub's or OpenAI's acts. Most of their claims are speculative and conclusory, not specific about the elements necessary to succeed on the merits.

OpenAI also moved to dismiss because the Does have not identified themselves. Courts do not usually allow plaintiffs to sue anonymously or pseudonymously absent special circumstances (for example, when there is a risk of retaliation). Procedural rules require plaintiffs to ask a court for permission to file lawsuits as Does. These Does failed to do this. It is, moreover, difficult for defendants to formulate adequate defenses if they do not know who is suing them.

Removal of Copyright Information Claims

The big money claim in the *Doe v. GitHub* lawsuit ($9 billion) asserts GitHub and OpenAI illegally removed CMI from source code used as training data.

Section 1202(c) defines CMI as including information identifying the work, its author and/or copyright owner, terms and conditions for use of the work, and/or identifying numbers or symbols representing identifying information.

To violate § 1202, a defendant must have intentionally removed CMI from copies of a work or must have distributed copies of a work knowing its CMI had been removed. In addition, a defendant must know or have reason to know that the CMI removal "will induce, enable, facilitate, or conceal an infringement" of copyright in the work.

Courts can award anywhere between $2,500 and $25,000 in statutory damages for each violation of this law. (When actual damages are difficult to prove, as with removal of CMI, legislatures sometimes decide to establish a statutory damage remedy to ensure some meaningful compensation is available to victims of a law's violation.)

This double knowledge requirement may be difficult to satisfy in the *Doe v. GitHub* lawsuit, as it was in *Stevens v. Corelogic*. Stevens is a photographer who specializes in taking digital photographs of houses on behalf of real estate agents. Stevens' photographs are typically posted on Multiple Listing Service (MLS) platforms.

Lawyers initiate lawsuits by filing complaints explaining the legal theories on which the lawsuits are based and core facts that support those legal theories.

Some metadata about Stevens' photographs is automatically created when his digital camera takes photographs. He can also add metadata to digital image files manually using photo-editing software. Metadata embedded in digital files may be invisible to anyone who looks at the image.

Corelogic provides software to MLS for displaying real estate photographs of houses for sale. Because image files can be very large, Corelogic resizes the images and saves the resized images so they occupy less storage space and load faster on MLS sites. In the process of resizing photographs, Corelogic's software did not preserve invisible metadata.

Stevens sued Corelogic for violating § 1202 because of its removal of CMI embedded in his photographs. The Ninth Circuit Court of Appeals affirmed a lower court ruling in Corelogic's favor, holding that Stevens had not shown that Corelogic intentionally removed the CMI, nor that the removal would facilitate copyright infringement.

GitHub and OpenAI argue the *Stevens* case supports their assertion the Does have not stated a viable claim for violation of § 1202.

Breach of License Claims

The Does' complaint identifies some open source licenses the Does themselves have used for software they developed that they claim GitHub and OpenAI have wrongfully included in Codex and Copilot. The Does say these and other class members' open source licenses require attribution and inclu-

sion of copyright notices in any reuses of their software.

As a defense against the breach of license claims, GitHub relies both on its terms of service and on license rights developers give GitHub when they choose to make their program code part of a public repository.

GitHub requires all of its users to agree to its terms of service. Included in these terms is a license granting GitHub the right to "store, archive, parse, and display ... and make incidental copies" of the users' code, as well as to "parse it into a search index or otherwise analyze it" and "share" the resulting code in public repositories with other users.

And when GitHub users decide to make their code available on its site, they must choose whether to make their code repositories private or public. Users who decide to make their repositories public "grant each User of GitHub a nonexclusive, worldwide license to use, display, and perform Your Content through the GitHub Service and to reproduce Your Content solely on GitHub as permitted through GitHub's functionality."

Why No Copyright Claim?

The biggest mystery in the *Doe v. GitHub* case is why there is no copyright claim in the complaint. One possibility is the Does are seeking copyright registration certificates for their programs. This is a necessary procedural requirement for U.S. copyright owners who want to sue someone for infringement.

Another possibility is the Does do not want to litigate fair use defenses that GitHub and OpenAI would almost certainly raise if sued for copyright infringement. (Fair use may not be a viable defense to the CMI removal or license breach claims.)

Fair uses are not infringements of copyrights. Courts consider four factors in making fair use determinations: the purpose of the challenged use; the nature of the copyrighted work; the amount and substantiality of the taking; and harms to the market for the work.

Existing U.S. precedents seem to support such a defense if the Does sue GitHub and OpenAI for infringement. The closest precedent is the *Authors Guild v. Google* case. The Second Circuit Court of Appeals held that Google had made fair use of millions of in-copyright books it scanned to enable computational analysis of a database of these books and for purposes of indexing their contents to serve up snippets of text in response to user search queries.

The court held that Google had made transformative uses of the in-copyright books because the corpus facilitated greater access to information. While Google copied the whole of each book, this was necessary to achieve its transformative purpose of indexing book contents for computational analysis and search. Because Google only served up three short snippets from each book, the snippets were unlikely to undercut the market for the books.

Under the *Google* decision, ingesting publicly available source code would seem to be as fair as the scanning of books to index their contents. And the snippets of code that Copilot provides in response to user prompts is analogous to the snippets of text from books that Google provides in response to user search queries. Because the court found both the scans and the snippets to be fair uses, GitHub and OpenAI would seem to have plausible fair use defenses.

Conclusion

Generative AI has raised some new technology issues courts have not yet addressed. While the *Doe v. GitHub* complaint raises some interesting theories of liability, it is far from clear courts will find Copilot or Codex to be unlawful. GitHub is arguing Copilot is socially beneficial because it "crystallizes the knowledge gained from billions of lines of public code, harnessing the collective power of open source software and putting it at every developer's fingertips." In May 2023, a trial court denied the GitHub and OpenAI motions to dismiss as to the removal of CMI and breach of license claims, so the lawsuit will now proceed to address the merits. It remains to be seen how receptive the court will be to GitHub's and OpenAI's defenses. C

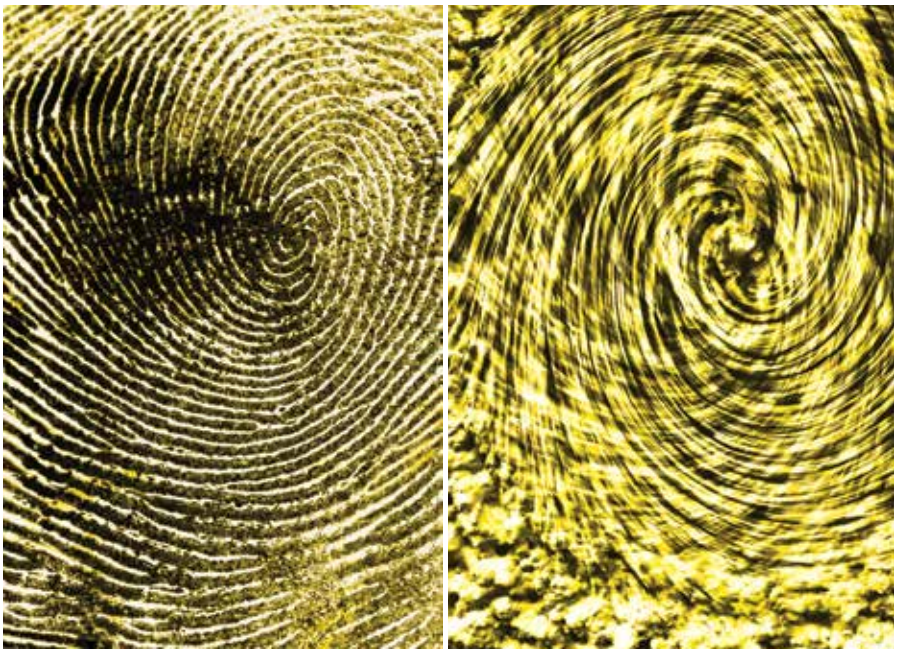
Pamela Samuelson (pam@law.berkeley.edu) is the Richard M. Sherman Distinguished Professor of Law at the University of California, Berkeley, CA, USA.

DOI:10.1145/3597173

Alessandro Acquisti and Ryan Steed

Privacy
Learning to Live with Privacy-Preserving Analytics

Seeking to close the gap between research and real-world applications of PPAs.



Over the past few decades, computer scientists and statisticians have developed tools to achieve the dual goals of protecting individuals' private data while permitting beneficial analysis of their data. Examples include techniques and standards such as blind signatures, *k*-anonymity, differential privacy, and federated learning. We refer to such approaches as *privacy-preserving analytics*, or PPAs. The privacy research community has grown increasingly interested in these tools. Their deployment, however, has also been met with controversy. The U.S. Census Bureau, for instance, has faced a lawsuit over its differentially private disclosure avoidance system; opposition to the new privacy plans garnered support from both politicians and civil rights groups.[3,11]

In theory, PPAs can offer a compromise between user privacy and statistical utility by helping researchers and organizations maneuver through the trade-offs between disclosure risks and data utility. In practice, the effects of these techniques are complex, obfuscated, and largely untested. While the interest in PPAs has been growing particularly among computer scientists and statisticians, there is a need for complementary social science research on the downstream impacts of these tools[13]—that is, the concrete ripples those technologies may cast for individuals and societies. In this Viewpoint, we advocate for an interdisciplinary, empirically grounded research agenda on PPAs that connects social and computer scientists.

A complete survey of this area of research would vastly exceed the space limitations for this Viewpoint.[a] Instead, after briefly summarizing what PPAs are, we focus on discussing how their rising deployment in real-world applications has been a cause of controversy, but also why PPAs are promising tools that both deserve and necessitate interdisciplinary research attention.

a For an annotated bibliography of additional references not included in this Viewpoint, see https://bit.ly/3I2lGHN

What Are Privacy-Preserving Analytics?

The term "privacy-enhancing technologies" (PETs) has been used to refer, broadly, to methods, tools, or technologies devised to protect individual privacy. By the term "privacy-preserving analytics," we refer to a particular subset of PETs that attempt to protect individual data while permitting some degree of data analytics.

Some PPAs are defined by specific methods for ensuring private analysis. Differential privacy, for example, can be applied to certain kinds of federated learning—a class of machine learning methods in which a model is trained across multiple devices without send-

ing the original training data to a centralized location—as well as to stochastic gradient descent (a common optimization technique in deep learning). Homomorphic encryption—another PPA on which applications such as blind signatures are based—permits computations (including predictive analytics) performed entirely on encrypted data without access to the decrypted version of the data. Similarly, the field of multiparty computation (MPC) is devoted to protocols that allow multiple parties to participate in aggregate computations without revealing their private data.

PPAs can also be described by the privacy definitions they satisfy. Standards such as *k*-anonymity—which requires an individual's attributes match no less than *k* other records in a dataset—or *l*-diversity are commonly used to evaluate the efficacy of particular PPA approaches, though these standards have since been shown to be vulnerable to certain re-identification attacks. For future-proof, formal guarantees, PPAs may be required to conform to differential privacy, which bounds the influence of any individual's data on output statistics. These standards are often achieved using the methods in the previous paragraph—for example, Google's abandoned Federated Learning of Cohorts (FLoC) proposed to provide *k*-anonymity through federated clustering.[8] Differential privacy can also be federated to provide local differential privacy, where each device applies a differentially private mechanism before aggregation.

Growing Applications

While several government organizations have been using, for a long time, an array of traditional statistical disclosure limitation methodologies, in recent years industry and government deployments of novel PPAs have become more common. Perhaps the most high-profile example is the U.S. Census Bureau's new Disclosure Avoidance System (DAS), which provides differential privacy guarantees for the 2020 Decennial Census. A team of bureau scientists realized the need for stronger protections after internal research suggested that published census data could be linked with commercial data to re-identify more than 52 million individuals.[3] The bureau had previously leveraged differential privacy to release additional, previously unavailable data through the 2008 On-The-Map tool.

PPAs are promising tools that both deserve and necessitate interdisciplinary research attention.

Another prominent example is Google's proposal to replace third-party cookies with interest-based alternatives. Using a federated approach, Google initially proposed to automatically group users into *k*-anonymous cohorts based on their personal data. More recently, Google proposed Topics API, a taxonomized approach to grouping users into much larger interest categories.[8]

PPAs are often proposed as a means to facilitate data sharing with researchers. In partnership with Social Science One, Facebook released in 2020 a large, differentially private set of URLs shared on the platform to aid studies on social media and democracy.[9] Google, LinkedIn, and Microsoft have also released public datasets with differential privacy.[6] Researchers have also extensively discussed and debated the use of *k*-anonymity to protect individuals' health data and satisfy the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[10]

Less publicized applications also exist. Differential privacy has been applied in deployed systems including the Facebook advertising stack: various internal SQL systems at Google, Uber, Oracle, SAP, and other tech companies; and local telemetry in Apple devices.[6] Federated learning is even more widespread, deployed in notable products such as Google's GBoard keyboard.

Growing Dissent

Our Viewpoint is motivated by the observation that, as PPAs have grown in

Many of the implementations of PPAs have raised controversy *precisely* because they engender a trade-off between statistical utility and data privacy.

popularity in research circles and as their real-world applications proliferate, their deployment has also been met by controversy. For instance, the introduction of differential privacy into Census Bureau operations has been particularly controversial, with a number of scholars, politicians, and civil rights groups questioning its use. Google delayed, and then shelved its initial third-party cookie replacement plans after privacy advocates protested, asserting it would make fingerprinting and discriminatory advertising easier.[5] Social Science One publicly disagreed with Facebook's interpretation that differential privacy was required for the Facebook URLs dataset, arguing that merely de-identified or aggregated academic data sharing should be enough to satisfy the General Data Protection Regulation and Facebook's FTC consent decree.[9] And researchers have questioned the use of privacy-preserving techniques in high-stakes settings (such as healthcare), fearing the utility trade-offs may especially affect vulnerable minority groups for whom PPAs may not perform as well.

We believe many of the implementations of PPAs have raised controversy *precisely* because they engender a trade-off between statistical utility and data privacy. In the census differential privacy example, demographers and politicians protested the deleterious effects that injected noise might have on critical research and political processes. In the FLoC example, privacy advocates protested the opposite—that Google sacrificed user privacy in order to preserve advertisers' ability to precisely target consumers.[5]

Granted, many studies in computer science and statistics already analyze these trade-offs. However, and key to our point, most approach the issue from formal perspectives, analyzing the trade-offs between abstract privacy parameters (for example, epsilon, for differential privacy) and abstracted loss functions (for example, root mean squared error).[1] In practice, the utility trade-offs reference concrete, not abstract, downstream benefits and harms; politicians are concerned that their electoral maps will be overhauled to the advantage of their opponents, and advertisers are concerned that the new Topics API will stifle sales.[12] Privacy trade-offs, on the other hand, often live in the realm of risk. Data stewards (as in the case of the census) must assess the chance of a potentially catastrophic reconstruction attack, while translating complex, probabilistic privacy parameters into concrete confidentiality guarantees.[11] Stakeholders may have practical preferences and concerns over these outcomes. However, the mostly theoretical literature makes it difficult to assess the ability of PPAs to satisfy all parties. We argue there is a need for research emphasizing empirical methods and participatory approaches involving a diverse set of stakeholders focusing on downstream outcomes and a more holistic set of consequences.

Downstream Implications, and Looking at the Bigger Picture

By downstream outcomes, we refer to empirical studies of the organizational and managerial considerations behind PPA development, along with the impact of their usage on those organizations in consumer products, in research, and in policymaking.

For example, in recent work,[15] we considered the controversy over the introduction of differential privacy to the 2020 Census and looked at potential but practical (and empirically measurable) implications of a deployment of differentially private algorithms in policymaking. Every year, Census estimates are used to guide the allocation of over $1.4 trillion in federal funding, including more than $16 billion dollars

in education funding divided among school districts in 2019. The allocation of that funding relies on a formula based on the Census Bureau's estimate of the number of children in poverty across the country. We simulated the effects of two different kinds of statistical uncertainty on the allocation of these Title I grants: the effect of noise injected to achieve differential privacy, and the effect of *existing* data error in the Census Bureau's estimates. What we found was that, while enforcing differential privacy did cause "misallocation" of funding relative to the official allocations, the misallocations caused by simulated data error were much larger. In other words, our results indicated the decision to augment privacy only adds to much larger costs of statistical uncertainty in census-guided grant programs. Weakening privacy protections would do little to mitigate these deeper disparities—in fact, if respondents are not confident in the privacy of their responses, the usefulness of census data will be further reduced, especially for harder-to-count groups.

In our study, the addition of differential privacy exposed deeper flaws in the way census-guided funding programs distribute the impacts of statistical uncertainty. But we also found that simple policy changes can reduce these impacts, ameliorating the costs of existing uncertainty and making the addition of stronger privacy protections less daunting. Our study shows the trade-offs involving PPAs may not be as stark as they seem: When PPAs conflict with existing data infrastructures, scientists and policymakers have an opportunity to revisit and improve statistical practices. Indeed, Oberski and Kreuter[13] argue that the constraints imposed by differential privacy (on repeated and granular analysis) could act as a barrier to *p*-hacking and other dodgy techniques.

We can already see glimpses of more robust, privacy-preserving research in studies of prominent, differentially private datasets. Evans and King[7] adapt existing methods for correcting naturally occurring data error to help construct valid linear regression estimates and descriptive statistics and evaluate their corrections on the differentially private Facebook URLs Dataset released in 2020.[9] Buntain et al.[4] develop a robust measure of ideological position on the same dataset and conduct an extensive empirical study of news platforms and audience engagement. In many cases, these more robust estimators significantly reduce the costs of privacy. For example, Agarwal and Singh[2] propose methods for noise-adjusted causal inference, recovering the results of a seminal study on import competition with differentially private 2020 Decennial Census data. And PPAs may require other, less technical changes with beneficial side effects—for example, interviews by Sarathy et al.[14] with differential privacy practitioners suggest a need for better data documentation and more context-specific education for data analysts.

What Is Next for PPAs?

In theory, PPAs can offer a compromise between user privacy and statistical utility. In practice, the effects of these techniques are still to a large extent untested. PPAs still pose challenges for scientists and policymakers. Work on correcting methods to account for noise and other privacy protections is still ongoing, and not all costs can be mitigated by more robust statistical practices. For example, studies by Sarathy et al.[14] and others describe practitioners' struggles with understanding and successfully applying PPA tools. And not all applications may be suitable for PPAs: some uses may be more suited to more traditional legal protections, and some uses may entrench harmful data practices. For example, the EFF protested Google's third-party cookie replacement not only because it could aid browser fingerprinting, but also because any form of targeted advertising can be used for exploitation and discrimination.[5] On the other hand, some uses of PPAs could help prevent harms like these: for example, PPAs could help facilitate sensitive data sharing with auditors, helping hold online platforms accountable for discrimination and other harms.

The controversies surrounding the deployment of PPAs highlight a critical gap in current research: While the interest in PPAs has been growing particularly among computer scientists and statisticians, there is still a paucity of applied, empirical research on the downstream implications of the development and deployment of PPAs across a variety of usage cases. Thus, there is a need for complementary social-science research on the ripples those technologies may cast—that is, the impacts of these tools. In the past couple of years, more applied research has started appearing. Our hope is that computer and social scientists will increasingly collaborate on research in this area to build generalizable knowledge and develop a grounded theory of the trade-offs involved, thus highlighting both the benefits as well as the burdens of PPA adoption for various stakeholders in real-world settings. C

References

1. Abowd, M. and Schmutte, I.M. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review 109*, 1 (2019), 171–202.
2. Agarwal, A. and Singh, R. Causal Inference with Corrupted Data: Measurement Error, Missing Values, Discretization, and Differential Privacy. (2022); https://bit.ly/3I2wi9j.
3. Bouk, D. and boyd, d. Democracy's Data Infrastructure. Knight First Amendment Institute. https://bit.ly/3pwPANM.
4. Buntain, C. Measuring the Ideology of Audiences for Web Links and Domains Using Differentially Private Engagement Data. (2021); https://bit.ly/41oCyzj
5. Cyphers, B. Google's FLoC Is a Terrible Idea. Electronic Frontier Foundation. (Oct. 2021); https://bit.ly/42LiWX9.
6. Desfontaines, D. A list of real-world uses of differential privacy. Ted is writing things. (2021); https://bit.ly/3BkS8kW.
7. Evans, G. and King, G. Statistically valid inferences from differentially private data releases, with application to the Facebook URLs dataset. *Political Analysis 31*, 1 (Jan. 2023); https://bit.ly/44PEtQj.
8. Goel, V. Get to know the new Topics API for Privacy Sandbox. Google. (May 16, 2022); https://bit.ly/3Bjk34t.
9. King, G. and Persily, N. Unprecedented Facebook URLs dataset now available for academic research through Social Science One. Social Science One. (202)); https://bit.ly/3MlOCwZ.
10. Malin, B. et al. Never too old for anonymity: A statistical standard for demographic data sharing via the HIPAA Privacy Rule. *Journal of the American Medical Informatics Association 18*, 1 (Jan. 2011); https://bit.ly/42tQyZU.
11. Nanayakkara, P. and Hullman, J. What's driving conflicts around differential privacy for the U.S. Census. *IEEE Secur. Privacy* (2022); https://bit.ly/42RKU3D.
12. Nguyen, G. Google's Topics API: Advertisers share concerns about topic diversity and other potential challenges. Search Engine Land. (2022); https://bit.ly/3VXKupP.
13. Oberski. D.L. and Kreuter, F. Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review 2*, 1 (Jan. 2020); https://bit.ly/44PQvcm.
14. Sarathy, J. et al. Don't Look at the Data! How Differential Privacy Reconfigures the Practices of Data Science. (2023); https://bit.ly/3BkOmb0.
15. Steed, R. et al. Policy impacts of statistical uncertainty and privacy. *Science 377*, 6609 (Aug. 2022), 928–931; https://bit.ly/3BgQXTB.

Alessandro Acquisti (acquisti@cmu.edu) is the Trustees Professor of Information Technology and Public Policy at the Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA.

Ryan Steed (ryansteed@cmu.edu) is a Ph.D. student at the Heinz College, Carnegie Mellon University, Pittsburgh, PA, USA.

The authors thank Priyanka Nanayakkara and Roy Rinberg for feedback on an earlier version of this Viewpoint. Alessandro Acquisti gratefully acknowledges support from the Alfred P. Sloan Foundation.

DOI:10.1145/3571280

Jeff A. Johnson and Daniel H. Bullock

Viewpoint

Fragility in AIs Using Artificial Neural Networks

Suggesting ways to reduce fragility in AI systems that include artificial neural networks.

ARTIFICIAL NEURAL NETWORKS (ANNs) are a promising technology for supporting decision making and autonomous behavior. However, most current-day ANNs suffer from a fault that hinders them from achieving their promise: fragility. Fragile AIs can easily make faulty recommendations and decisions, and even execute faulty behavior, so reducing ANN fragility is highly desirable. In this Viewpoint, we describe issues involved in resolving ANN fragility and possible ways to do so. Our analysis is based on what is known about natural neural networks (NNNs), that is, animal nervous systems, as well as what is known about ANNs.

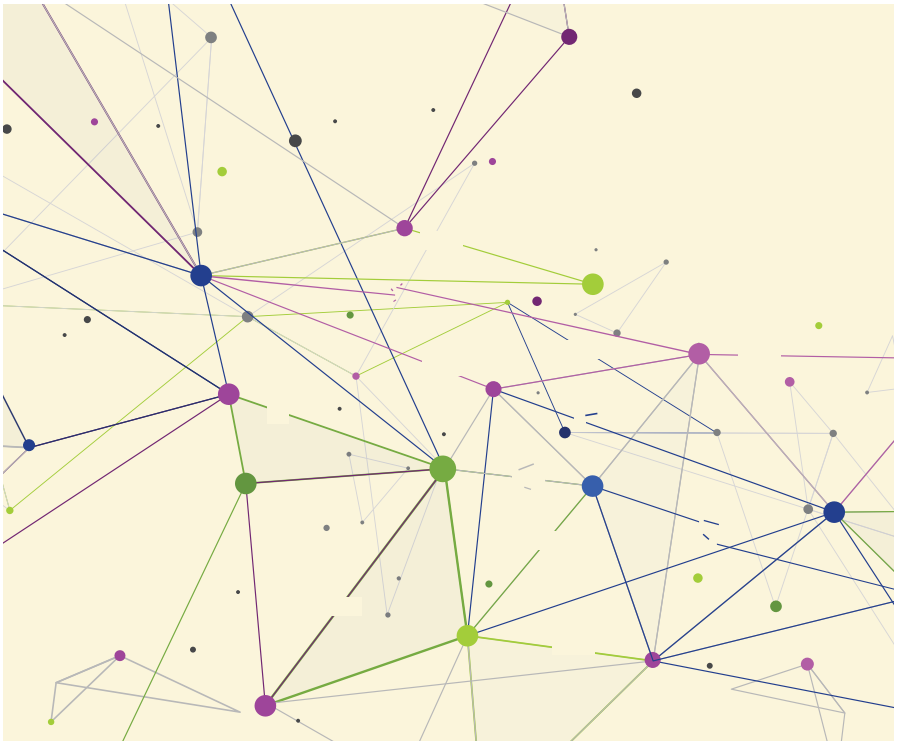


IMAGE BY RAMCREATIVE

What Is Fragility?

Engineered systems fall on a continuum from robust to fragile. One reason some ANNs are considered "fragile" is that seemingly minor changes in the data they are given can cause major shifts in how they classify the data. Such fragility is often evident when lab-trained ANNs are finally tested under real-world conditions. A classic example of hyper-fragility is an ANN that, after being trained to recognize images of stop signs, fails to recognize ones in which a small percentage of pixels has been altered.

Inputs: Passively received vs. actively gathered. First, it must be recognized that fragility is not unique to ANNs. NNNs, for example, animal brains, can also be confused by minor alterations of inputs. All animals can be tricked by well-crafted or naturally arising "lures." Synthesized optical illusions are a prime example. In Figure 1, minor shifts in the relative positions of light and dark areas within two nearby bands cause the human visual system to see either spirals (left) or concentric circles (right).

However, unlike most ANNs, human sensory systems ride on mobile segments of an actor. Shown the illusion in Figure 1, many people vary their angle of regard and scan pattern, and some may trace a finger along one of the "spirals," and thereby discover they have two unlinked circles, not a spiral. Similarly, people and other animals often do "double-takes" when uncertain or when they perceive something novel or unexpected given a context. In so doing, they are actively gathering additional data to compensate for their visual system's fragility.

Stated more generally, when constrained to be completely passive, NNNs can be fragile, but when they output a categorization with a low confidence level, the containing executive system typically attempts to

gather more data. In contrast, most ANNs are implemented as purely passive receivers of data, with no containing executive, so we should not be surprised they often exhibit fragility. This suggests one powerful way to reduce fragility in ANNs: wrap them in executive systems that trigger active ("Gibsonian") data gathering[8] under specified conditions.

More sophisticated systems would detect more of these types of conditions. Perhaps the most general, and easiest to exploit, are confidence levels of the ANN's classifications: The executive can use low confidence levels to trigger second takes. Because ANNs are based on numerical scoring, it is straightforward to add a confidence metric and a threshold level, below which the AI would engage active information gathering. In other words, to reduce fragility, AIs need at least the ability to measure the degree of certainty of their perceptual decisions and, when certainty is low, do double-takes, for example, activate usually immobile or dormant sensors, or secondary data sources. Nonfragile AIs, by design, will withhold action whenever uncertainty cannot be resolved.

Confidence levels can be measured with purely feedforward ANNs, by computing an answer to this question: How strong, relative to correct past classification episodes, is the current evidence for applying this label? But more robustness could be achieved in ANNs augmented by feedback, which could also compute a certainty-relevant metric to answer the inverse question: Given this provisional label, how well does the current input dataset match the typical input that activated this label on successful past classifications?

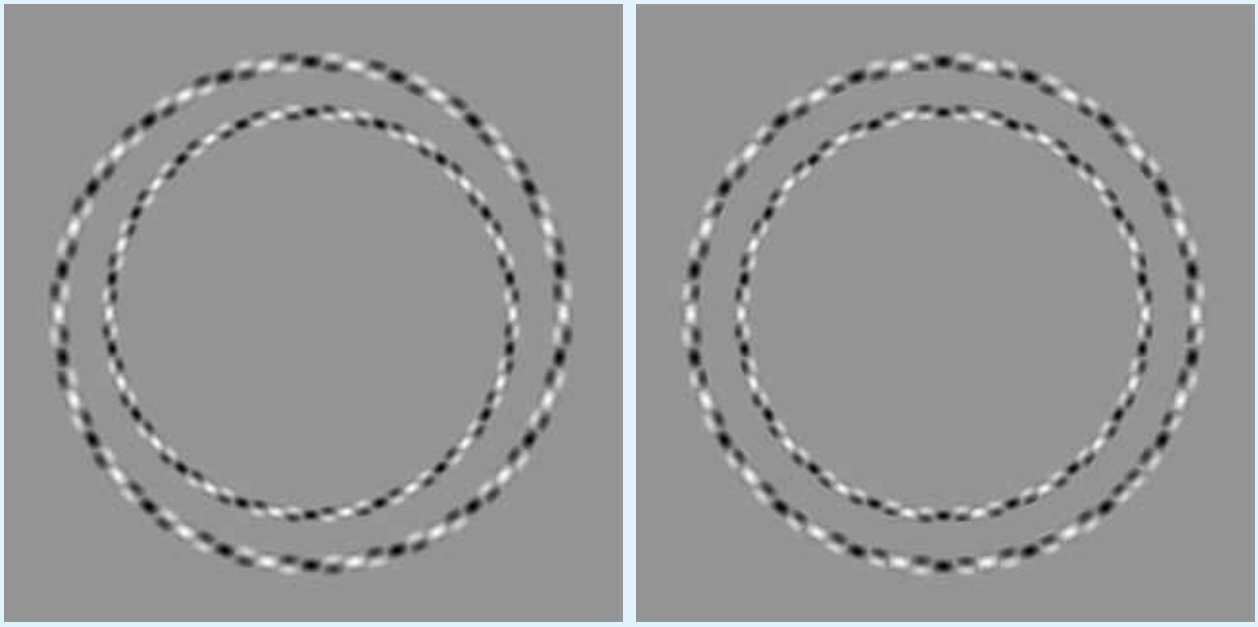
Engineered systems fall on a continuum from robust to fragile.

Autonomous driver AIs or aerial drone AIs are inherently equipped to improve greatly upon passive information processing of passive data streams from their sensor suites, because they move. Movement creates image sequences, providing rich information for disambiguation, but further robustness requires making the movements responsive to the confidence/uncertainty metrics produced within the ANNs.

Note also that the uncertainty threshold for engaging double-takes can be adaptive. It can be reduced in contexts where lures are unlikely, but increased in contexts where lures are a common threat to accurate classifications. This feature of an AI would allow it to mimic the flexible way that humans trade off speed (little or no pausing for double takes) and accuracy (ample time devoted to active information gathering). There is now impressive evidence that uncertainty causes humans to switch from a fast habitual mode of decision making to a slower, more deliberative mode.[1] This is a hallmark of executive control, and nonfragile AIs must mimic humans in this regard.

Single NN vs. multiple diverse NNs. A second cause of ANN fragility is that most consist of a single network comprising input (sensor-driven) neurons, inner (hidden) layers of neurons, and output neurons (indicators or effectors). Even if an ANN is deep in the sense that it has several inner layers of neurons, it is nonetheless a single ANN with a uniform structure and uniformly applied methods for adapting connections between neurons and for processing signal streams. For example, assume we have trained a "deep learning" ANN to recognize dog breeds depicted in photos. That ANN

Figure 1. Left: Concentric circles appear as spirals. Trace one with a finger. Right: Same figure with every 4th small rectangle flipped eliminates the illusion of spirals.



may be fragile, that is, it can possibly be confused by shadows, partial occlusion, alteration of a few carefully chosen pixels, or other minor fluctuations in the input.

In contrast, real animal (including human) brains and perceptual systems are not a single *N*; they consist of many interconnected NNs that differ widely in anatomy and physiology, that is, they may have different neural structures and use distinct learning processes affected by different combinations of modulating signals, for example, transient neurotransmitter release events triggered by prediction errors or uncertainty. These different neural networks were initially added to the nervous system by independent mutation-selection cycles during different evolutionary periods, and have been further refined in descendant lineages. They may operate in parallel or serially.

For example, anatomical and physiological studies indicate that primate visual systems are highly modular, with several distinct NNs processing the same visual inputs differently and combining their outputs to create the phenomenon of vision[9,10] and even blind-sight (the ability to utilize visible information with no subjective awareness of seeing). For example, thalamus, superior colliculus, and a dozen visual cortices all mediate perceptual decisions driven by visual inputs. Similarly, in humans and other mammals, the hippocampus, amygdala, and cortex—quite different brain structures—are all capable of mediating memory retrieval,[2,3] and memory-guided decisions.

Animals' cognitive architecture reduces fragility by consisting of multiple very different NNs evaluating the same stimulus, with each coming to a decision, plus mechanisms for integrating the outputs from those various NNs to produce an overall decision. Thus, outputs from distinct NNs can cooperate, but also compete, to produce a final decision or behavior. The fragility of the combined identification is reduced via weighted voting by non-equivalent decision agents. As in the colloquial expression, "two brains are better than one," it is also true that one brain with all anatomically distinct information processing streams intact is much better than a brain reduced to one homogeneous NN. This relates back to the active perception theme: binocular vision reduces visual ambiguity, relative to monocular vision, and an animal reduced by injury to monocular vision can compensate by moving its angle of regard, then reaching a decision based on an aggregation of multiple, spatially disparate, images.[8]

Some component NNs in a perceptual system produce results faster than others, which may render decisions by the slower component NNs moot. For example, if a large object comes flying at your head, some NNs in your brain will cause you to flinch, duck, or throw up your hands to block it, long before other NNs in your brain identify the object.[3,5]

AIs need at least the ability to measure the degree of certainty of their perceptual decisions and, when certainty is low, do double-takes.

Figure 2. Image of demonstration posted on social media (anonymized).

Consider the following real example of human perceptual fragility. In spring 2020, there were demonstrations across the U.S. against police killings of Black people. Most of the demonstrations were peaceful, but some were not. Many people posted photos and videos of the demonstrations on social media. One such photo (see Figure 2) shows people walking on a street. The person who posted the photo expressed horror that the man on the right was carrying an assault rifle, and asked why police—who were present—did not arrest him, especially since there had been shootings at other recent demonstrations. Several commenters—some of whom were at that demonstration—pointed out it was a rainy day (note the open umbrellas in the background) and suggested that what the man was carrying was just a closed umbrella. The original poster responded: "Yes, of course! Duh!" and wondered why she had perceived a gun. She initially saw a gun because recent news had trained one of her NNs to expect guns at these demonstrations. Other people were less certain and withheld decision until affected by slower NNs that took the context—the weather and the fact that other protesters and the police were not reacting as one would expect to the presence of someone carrying an assault rifle—into account, suppressing the alternative "take" that the black elongated shape was a rifle.

In a decision system comprising multiple component NNs, the relationship between the components—how they interact to produce an overall decision—is very important, and can be flexible. For example, the relative decision speeds of diverse component NNs influence how they interact: faster NNs would normally "win." Similarly, if the overall system weights the outputs of its component NNs differently, that would also influence the outcome. But humans and other animals can learn to exercise "executive control" in which they context-dependently inhibit the faster processes if they repeatedly produce faulty results in a context. This ability is highlighted in folk wisdom: "Fool me once, shame on you; fool me twice, shame on me."

Perceptual classification vs. proof in practice. The point just made leads to a much broader conclusion. An underappreciated source of fragility is that in typical ANNs, the loss function pertains only to whether the inputs are being used to make the best assignment of the current stimulus to classes/labels. In the real brain, there is an additional step involving reinforcement learning (RL): the final arbiter is whether a classification can effectively guide behavior that leads to reward.[4,6] Earlier, we emphasized classification uncertainty, because the best match may not be a good enough match. But there are really two senses of good match. The sense already highlighted is purely cognitive: Is the current exemplar similar enough to the template it best matches to warrant a high expectation that it is a true member of the category?

A second sense is purely practical: Does the behavior triggered by making this category assignment actually pay off? To exemplify, imagine a shop keeper presented with a $50 bill from customer X. He cannot see any way that it departs from his mental template of a genuine $50 bill: His perceptual assessment yields high confidence. But after he deposits it at his bank, he is notified that it was rejected as counterfeit. Via reinforcement learning, repeated episodes of this type will cause the shopkeeper to stop trusting his visual assessment of $50 bills presented to him by customer X. In this way, RL teaches an actor that the best its cognitive system can do, in a particular context, is not good enough to warrant a given behavior.

There are really two senses of good match.

This point dovetails with the need for multiple ANNs that make independent contributions. In the human visual system and hippocampal memory system, separate streams exist to represent spatial variables/contexts versus the cues/objects that occupy spatially segregated contexts.[7] Because the behavioral implications of cues and objects are so context-dependent, such separate streams are vital for implementing the RL strategy just noted.

The foregoing analysis suggests further practical ways to avoid fragility in ANNs:

- Combine several ANNs with very different structures, operating characteristics, speeds, and internal connection-weights.
- One plausible group-structure is cascading systems in which inexpensive, fast systems make rough assessments of the data and slower, more precise systems make successively more refined assessments, until at some point the whole system issues a response.
- Include an input ANN that feeds all the decision ANNs the same inputs.
- Include an output executive that assigns varied weights to the outputs of the component ANNs and combines their results.

This strategy can be seen as following the principle that group decisions by voting are of higher quality, but only if the voters composing the group are diverse. The diversity is as important as the executive control strategy. Ideally, the distinct NNs should have uncorrelated failure modes. Mere ANN partial cloning, for example, training three ANNs of the same type from different initializing weight matrices, would increase "reliability through redundancy," but cannot greatly improve decision quality.

Conclusion

We have argued that fragility in ANNs can be reduced by three related means:

- Wrap them in executive systems that monitor ANN confidence levels and, when confidence is low, trigger processes to actively seek more information.
- Modularize ANN-based decision systems to consist of multiple ANNs with different structures and operating characteristics, with executive systems to combine the outputs of the component ANNs.
- Recognize the best classification of a perceptual subsystem may not be good enough to guide behavior in some contexts, and include a RL-trained ANN stage to provide a context-sensitive capacity to reversibly enable or disable each classification as a trigger for specific behaviors.

Some researchers are starting to use some of these methods to reduce fragility, but until these measures become widely adopted outside of academia, fragility in ANNs will continue to be a problem. C

References

1. Daw, N.D. et al. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience 8* (2005), 1704–1711.
2. Eagleman, D. *Incognito: The Secret Lives of the Brain.* Vintage Books, New York (2012).
3. Eagleman, D. *The Brain: The Story of You.* Vintage Press, New York (2015).
4. John, Y.J. et al. Anatomy and computational modeling of networks underlying cognitive-emotional interactions. *Frontiers in Human Neuroscience 7*, 101 (2013).
5. Kahneman, D. *Thinking Fast and Slow.* Farrar Straus and Giroux, New York (2011).
6. Patrick, S. and Bullock, D. Graded striatal learning parameters enable switches between goal-directed and habitual modes by reassigning behavior control to the fastest-computed reward predictive representation. bioRxiv, (2019); https://doi.org/10.1101/619445.
7. Ritchey, M. et al. Dissociable medial temporal pathways for encoding emotional item and context information. *Neuropsychologia 124* (2019), 66–78.
8. Rucci, M. et al. Integrating robotics and neuroscience: Brains for robots, bodies for brains. *Advanced Robotics 21* (2007), 1115–1129.
9. SfN/BrainFacts Vision: Processing Information, Brain Facts, Society for Neuroscience. (2012); https://bit.ly/3BrbMM0
10. Van Essen, D.C. *Information Processing in the Primate Visual System, Advances in the Modularity of Vision: Selections from a Symposium on Frontiers of Visual Science*, Washington, D.C., National Academies Press, (1990); https://bit.ly/3pK1DHT

Jeff A. Johnson (jajohnson9@usfca.edu) is a retired assistant professor in the computer science department at the University of San Francisco, CA, USA.

Daniel H. Bullock (danb@bu.edu) is Professor Emeritus in the Department of Psychological and Brain Sciences, Boston University, MA, USA.

DOI:10.1145/3574156

Tariq Zaman

Viewpoint

Lost in Afghanistan: Can the World Take ICT4D Seriously?

Seeking to maximize the value of information and communication technologies for development research.

In August 2021, I was reading Lt. Gen. Sami Sadat's article in the *NY Times*,[24] and it reminded me of the lyrics of Billy Hill's song: "I tried and I tried and I tried and I tried. To make them understand. I tried and I tried and I tried and I tried. But they just can't understand." Like any other Pashtun, I also consider Afghanistan my "Loy Kor" or "Greater Homeland." Therefore, the early days of Afghanistan's fall were full of terrible surprises. There are many more questions in my mind than answers; consequently, it is normal to reflect on the situation from my own academic and professional background in information and communication technologies for development (ICT4D) and computer-human interaction (CHI).

An Afghan policeman sits near private cellphone antenna in Kandahar, Afghanistan.

PHOTO BY JAVED TANVEER/AFP VIA GETTY IMAGES

ICT4D is an interdisciplinary field that has emerged in the 1990s when mobile phones, PCs, and Web services became cheaper and accessible in developing countries, and governments and international organizations heavily funded ICT-based initiatives as the tools for achieving development goals.[7] In the early days, ICT4D research and practices were influenced by the technology diffusion models and overlooked the integration of co-design and factors of social embeddedness of ICTs, and therefore failed to contribute toward the achievement of inclusive development.[16] However, the lessons learned are well documented and have led to developing specific guidelines and new watchwords.[4] Based on my academic and ethnic background and recent fieldwork in the northwest part of Pakistan, I present my reflections on the Afghanistan fiasco in this Viewpoint. Following the sutra of effective communication,[20] I restrict my reflections to three main points.

The imperative of a defined exit strategy. The U.S. did not have an exit strategy for withdrawal from Afghanistan. In December 2018, the Taliban announced their meeting with U.S. officials to find a "roadmap to peace" but refused to hold official talks with the Afghan government.[3] The U.S. government subsequently signed the Agreement for Bringing Peace to Afghanistan (2020) and made the situation worse by not making the Afghan government (their own ally) a party to the agreement.[29]

The CHI and ICT4D community has extensively highlighted the importance of a responsible exit strategy for avoiding software and hardware suppliers' hostage situations,[6,11,27] maintaining the secrecy and confidentiality of vulnerable populations,[13] establishing the

participants' willingness and capability for taking over the project,[9] and controlling the dependencies and expectations of the stakeholders.[12] Luke Jordan, in his recent guidebook (which I think is a must-read for anyone in the field of ICT4D), says, "a product that looks wildly different at the end than at the beginning could mean a terrible starting point."[21] Carl Gunnstam and Carl Johan Nordquist highlighted the need for an exit strategy in ICT4D interventions. They reported, "a well-designed exit plan that is agreed upon by all stakeholders diminishes the risk that the project is untimely abandoned by a stakeholder, and thereby increases the likelihood of a sustainable project."[15]

In 2016, during my field study in the forests of Malaysian Borneo, an indigenous Penan village headman explained to me some important aspects of "Smart Villages," including the need to engage partners from the early stages of the project's planning to discuss the long-term sustainability of the interventions.[31] While the Penans are more concerned about the health and safety implications of non-functioning solar-powered batteries of the telecenter project, an Afghan artist sent 20 tons of rubbish collected from Bagram airbase, labeling it as art and "A gift to the American people."[5]

Open source software (OSS) vs. proprietary software. It was shocking to learn that as soon as the U.S. troops started to withdraw, the U.S.-hired contractors took proprietary software and weapons systems with them.[14] By July 2021, most of the 17,000 contractors had left, and with them, access to the software which the Afghan Army relied on to track their vehicles, weapons, and personnel also disappeared. As a result, the whole food and ammunition supply system collapsed.

In January 2004, the Swedish International Development Cooperation Agency (SIDA) published a report on open source technologies in developing countries, highlighting the use of OSS to assert security and economic independence for developing countries.[30] The report states, "the imperative to adopt OSS in these countries, particularly in the public sector, is also motivated by a desire for independence, a drive for security and autonomy, and to address intellectual property rights enforcement."

The CHI and ICT4D community has extensively highlighted the importance of a responsible exit strategy for avoiding software and hardware suppliers' hostage situations.

Local solutions and frugal design. Sadat revealed the Taliban fought with snipers and improvised explosive devices while we lost aerial and laser-guided weapon capacity.[24] The Afghan forces' supply lines depended on private contractors and their tools. Therefore, when the private contractors suddenly departed, the Afghan soldiers lacked the necessary tools to fight and surrendered.[8]

During the Cold War, the U.S. and Mujahideen heavily relied on an indigenous arms manufacturing industry of Pakistani Tribal areas that has existed for hundreds of years.[18] However, during the War on Terror, the Taliban developed professional gunsmith skills to design and manufacture improvised weaponry from locally available material scrap, auto parts, and obsolete weapons.[2] On the other hand, the Afghan Army relied heavily on advanced technology, complex logistics, and airpower, which it could not sustain independently.

The first essential reading for my undergraduate students of human-computer interaction is Bonnie Nardi's article, "Designing for the Future: But Which One?"[23] Nardi emphasized the notion of a Steampunk future derived from a past in which people exerted personal control over material culture. Nardi also reports the values of the Steampunk movement, which are: recycling materials, finding other functions for objects, expanding horizons, and regaining control over the fabrication of our everyday objects. As they say, context is the king: I use Nardi's article to open a debate on relating the computing students' preferences of hard and soft skills; the importance of local, social, indigenous, and cultural context; civic engagement; and the societal challenges of past, present, and future.

In the last 70 years, the "third world" or "global south" has been bombarded with "development theories" and international development agencies to put these countries through a "Western route to development." However, they have failed to bring positive changes in the lives of the target populations.[1,17] In CHI and ICT4D fields, a decade ago, postcolonial computing was introduced[19] to examine and address the lack of cross-cultural sensitivity in the design "for the developing world." However, postcolonial theories have been chastised for failing to include marginalized viewpoints and are considered "Eurocentric critiques of Eurocentrism."[1] To break the "colonized mind-set," Schultz suggests embracing plurality in design and development by including the marginalized perspectives (that is, marginalized immigrants, indigenous communities and the global-south populations).[25] Therefore, the decolonial perspective in ICT4D strongly echoed notions of pluriversal design,[10,22] creating a multiplicity of voices and valuing local knowledge and inclusive participation. Western governments fund ICT4D research[26,28] and Western academics are heavily represented and among the most active participants in generating useful knowledge in ICT4D. Academic conferences and literature are widely expected to produce networking and knowledge exchange between academics, policymakers and industry players. However, there is only piecemeal evidence of ICT4D research informing policymaking and practices despite the presence of very loud voices and research pieces of evidence. The ICT4D researchers failed to understand the non-academic audiences' state of mind. One explanation for this failure is the cognitive bias, which the psychologist Steven Pinker has termed as "the curse of knowledge." Here, I provide two recommendations to maximize the value of ICT4D research. First, ICT4D researchers can

The decolonial perspective in ICT4D strongly echoed notions of pluriversal design, creating a multiplicity of voices and valuing local knowledge and inclusive participation.

curb the curse of knowledge by developing four critical skills:

- Being able to understand the political process and identify key stakeholders;
- Being able to synthesize research findings with simple, compelling stories;
- Being a good networker; and
- Being able to connect and align all these factors in an institutional setup.

Second, the objectives of ICT4D conferences should include deliberate retrospection of the learning experiences, critical analysis of the existing development practices, and shaping the post-research development agenda to demonstrate the practical benefit of the ICT4D research.

The case of the defeat in Afghanistan is the most recent example of not listening to the experiences of the research community and learning from past mistakes. It again reminds me of Billy Hill's: "I tried and I tried and I tried and I tried/ But they just can't understand/Poor God, people/When will they ever learn?" C

References

1. Ali, S.M. A brief introduction to decolonial computing. *XRDS 22*, 4 (2016), 16–21.
2. Allison, J. Innovation and the improvised explosive device (IED). In *Understanding Terrorism Innovation and Learning*, Routledge (2015), 53–75.
3. Behuria, A. US-Taliban talks for Afghan peace: Complexities galore. *Strategic Analysis 43*, 2 (2019), 126–137.
4. Bon, A. and Akkermans, H. Digital development: Elements of a critical ICT4D theory and praxis. In *Proceedings of the Intern.Conf. on Social Implications of Computers in Developing Countries*. Springer, Cham. (2019), 26–38.
5. Chaves, A. Artist sends 20 tonnes of rubbish from Afghanistan to U.S. as 'gift to the American people'. *The National*. (2022); https://bit.ly/438hZIn.
6. Damsgaard, J. and Karlsbjerg, J. Seven principles for selecting software packages. *Commun. ACM 53*, 8 (Aug. 2010), 63–71.
7. De, R. et al. ICT4D research: A call for a strong critical approach. *Information Technology for Development 24*, 1 (2018), 63–94.
8. Detsch, J. Departure of private contractors was a turning point in Afghanistan military collapse. *Foreign Policy*. (2021); https://bit.ly/3pV4UDZ
9. Dreessen, K. et al. Toward reciprocity in participatory design processes. In *Proceedings of the 16th Participatory Design Conference 2020-Participation (s) Otherwise-Volume 2* (June 2020), 154–158.
10. Escobar, A. *Designs for the Pluriverse: Radical Interdependence, Autonomy, and the Making of Worlds*. Duke University Press, Durham (2018).
11. Fiesser, A. Buying software: Focus group summary. In *Proceedings of the 21st European Conf. on Pattern Languages of Programs* (2016), 1–4.
12. Frauenberger, C. et al. In pursuit of rigour and accountability in participatory design. *Intern. J. of Human-Computer Studies 74* (2015), 93–106.
13. Furtado, T.V. and Rechena, A. MIRAGE–The social function of artistic practice as a tool for empowerment. Creative net art projects with women in shelters. (2021).
14. Gibbons-Neff, T. and Schmitt, E. Multi-sided Civil War could grip Afghanistan, U.S. general warns. *The New York Times*. (2021); https://nyti.ms/3pNsZN5
15. Gunnstam, C. and Nordqvist, C.J. *A Study of the Sustainability of ICT- Projects in Developing Countries*. (2019); https://bit.ly/42KrM7U
16. Harris, R.W. How ICT4D research fails the poor. *Information Technology for Development 22*, 1 (2016), 177–192.
17. Haq, K. and Uner, K. *Human Development: The Neglected Dimension*. United Nations. (1986).
18. Hilali, A.Z. The costs and benefits of the Afghan War for Pakistan. *Contemporary South Asia 11*, (2002), 291–310.
19. Irani, L. et al. Postcolonial computing: A lens on design and development. In *Proceedings of the SIGCHI Conf. on Human Factors in Computing Systems* (2010), 1311–1320.
20. Limoncelli, T.A. Communicate using the numbers 1, 2, 3, and more. *Commun. ACM 63*, 6 (June 2020), 42–44.
21. Luke, J. *Don't Build It: A Guide For Practitioners In Civic Tech/Tech For Development*, Grassroot (South Africa) and MIT Governance Lab (U.S.) (2021); https://bit.ly/41NYfsF.
22. Masset, E. et al. Does research reduce poverty? Assessing the welfare impacts of policy-oriented research in agriculture. *IDS Working Papers, 2011* (360 (2011), 1–64.
23. Nardi, B. Designing for the future: But which one? *Interactions 23*, 1 (Dec. 2015), 26–33.
24. Sadat, S. Opinion | The Afghan Army collapsed against the Taliban. Here's why. *The New York Times*. (Aug. 25, 2021); https://nyti.ms/3MBP5LF.
25. Schultz, T. et al. What is at stake with decolonizing design? A roundtable. *Design and Culture 10*, 1 (2018), 81–101.
26. SPIDER: Spidercenter.org, (2022); https://bit.ly/3MwlSS5.
27. Taylor, N. and Cheverst, K. Creating a rural community display with local engagement. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems* (2010); 218–227.
28. USAID: Technology, (2021); https://bit.ly/3BzhgV3.
29. Verma, R. The U.S.-Taliban peace deal and India's strategic options. *Australian Journal of International Affairs 75*, 1 (2021), 10–14.
30. Weerawarana, S. and Weeratunge, J. *Open Source in Developing Countries*. Sida, Stockholm. (2004); https://bit.ly/3IlyZ6h.
31. Zaman, T. *It Is Not a Village but People: Long Lamai, a Case Study of a Smart Village*. Smart Villages Research Group. (2016); https://bit.ly/3pJTO4N.

Tariq Zaman (zamantariq@gmail.com) is an associate professor in the School of Computing and Creative Media, and Head of the Advanced Centre for Sustainable Socio-Economic and Technological Development (ASSET) University of Technology Sarawak (UTS) Jalan Universiti, Sarawak, Malaysia.

I am thankful to Claudia Canales Holzeis, Heike Winschiers-Theophilus and Khairuddin Hamid for their helpful comments.

ACM Acknowledges the Companies Participating in ACM's Preferred Employer Program

The ACM Preferred Employer Program allows startups and small companies to provide ACM Professional Membership and ACM Digital Library access to their technical staffs at a preferred rate.

The following companies currently participate in ACM's Preferred Employer Program:

- Biteable
- Chameleon Consulting Group, LLC
- Circonus, Inc
- City of Portland - Bureau of Technology Services
- CompSci, LLC
- CrowdStrike
- Fastmail
- FullStory, Inc
- HashiCorp
- IntelliGenesis LLC
- Joyent, Inc.
- Lightelligence, Inc.
- Luminous Computing
- Micron Technology
- Nielsen Norman Group
- Riskfuel
- Surfshark
- tc1
- Van Andel Research Institute
- Virtek Vision International

Through the ACM Preferred Employer Program, each team member receives all the benefits of ACM Professional Membership, including *Communications of the ACM*, discounted registration fees for ACM SIG conferences, video-based online courses, digital books, skill-based learning paths, hands-on practice labs, coding sandboxes and more from leading providers Pluralsight Skills and Skillsoft Percipio.

For more information and to apply:
www.acm.org/membership/preferred-employer

acm Preferred Employer

acm Association for Computing Machinery

East Asia and Oceania Region Special Section

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

DOI:10.1145/3598580

Welcome

WELCOME TO THE special section highlighting cutting-edge research and innovation emerging from East Asia and Oceania. Our region encompasses Southeast Asia, Oceania, and Asia-Pacific countries, including Japan and Korea. The articles in this section—designated as "Hot Topics" and "Big Trends"—aim to not only showcase technological advancements from this region, but also to strengthen research collaboration and communication with regions worldwide.

This special section brings together some of the most innovative research in computer science and technology from this flourishing region. The articles cover a wide range of topics, from state-of-the-art developments in learning analytics, AI and machine learning, education, Big Data, neuromorphic computing, and blockchain technology, to applications in disease prediction and assistive devices. They demonstrate the groundbreaking work being done to push the boundaries of our understanding and capabilities in these fields.

This special section also includes many nationwide and major cross-institutional collaborative works, including Australian efforts tackling misinformation and operationalizing responsible AI at scale, South Korea's nationwide endeavors in the AI semiconductor industry, a Japan-wide undertaking to establish high-speed, high-density, and highly secure networks across the nation, as well as Japan's Society 5.0 progress for exhibition in Osaka in 2025, New Zealand's work addressing veracity grand challenges, and a pioneering effort by a Vietnamese startup to nurture AI development in that region. There are also a wide array of country- and culture-specific technical challenges and developments discussed in these articles, such as those in manufacturing and fabrication. Get ready to be inspired and amazed by the array of topics that have the potential to shape our future in exciting and meaningful ways.

To create this special section, we organized a virtual workshop on October 28, 2022, and launched an open call for contributions. The workshop concluded with the selection of articles for this special section, which consists of four "Big Trends" articles and 15 "Hot Topics."

We extend our gratitude to the speakers at the workshop and all the authors who contributed to this special section.

—Flora Salim, Bingsheng He, and Ken-ichi Kawarabayashi
Co-organizers of East Asia and Oceania Region Special Section

Flora Salim is a professor of Computer Science and Cisco Chair of Digital Transport in the School of Computer Science and Engineering, at the University of New South Wales (UNSW) in Sydney, Australia.

Bingsheng He is a professor and Vice Dean (Research) in the School of Computing at National University of Singapore.

Ken-ichi Kawarabayashi is a professor at the National Institute of Informatics in Tokyo, Japan.

Watch the co-organizers discuss this section in the exclusive *Communications* video. https://cacm.acm.org/videos/east-asia-and-oceania-region-2023

EAST ASIA & OCEANIA REGION

SPECIAL SECTION

Big Trends

IMAGES BY: (L) HURCA; (R) ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

Hot Topics

Association for Computing Machinery
Advancing Computing as a Science & Profession

DOI:10.1145/3588431

BY DAMIANO SPINA, MARK SANDERSON, DANIEL ANGUS, GIANLUCA DEMARTINI, DANA MCKAY, LAUREN L. SALING, AND RYEN W. WHITE

Human-AI Cooperation to Tackle Misinformation and Polarization

A DOMINANT NARRATIVE of the past decade is that algorithms contribute to a misinformed and segregated society. Perhaps paradoxically, algorithms are often sought as solutions to such problems. We describe a significant emerging trend away from this techno-solutionist approach that seeks to create and understand a new paradigm: a productive interplay between algorithms and people. Two relevant test cases are being explored in our region: The first addresses a new framework to tackle misinformation by assisting fact-checkers with computational methods, and the second seeks new models to understand how search engines deliver personalized search results when little or no algorithmic personalization exists.

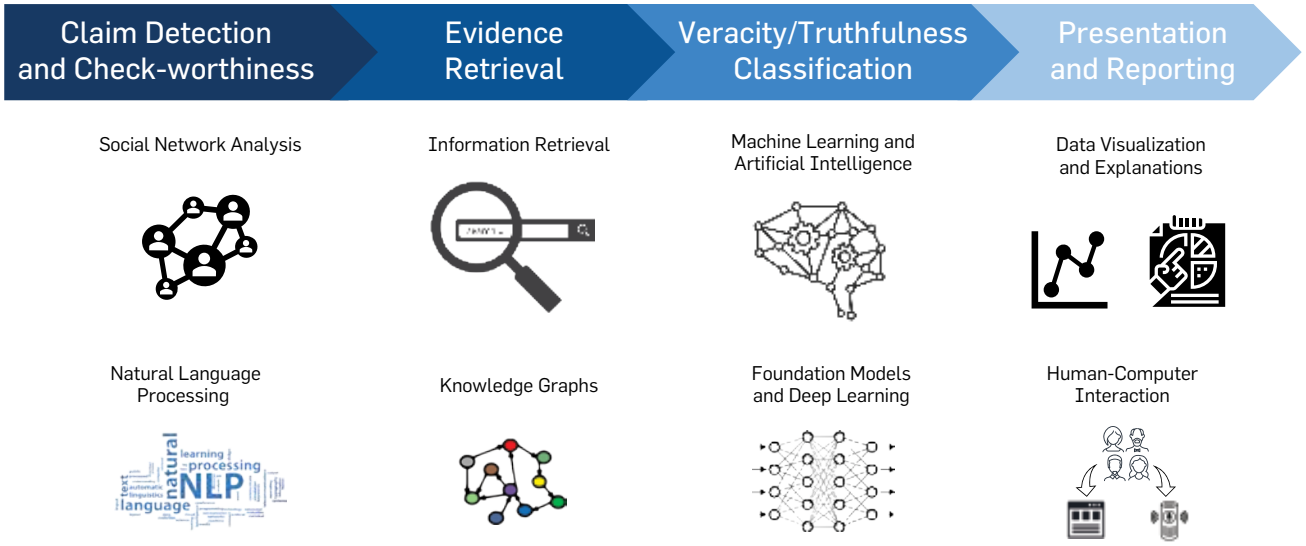
In late 2020 and early 2021, the Australian Communication and Media Authority conducted a study to analyze the state of misinformation in Australia. The findings, reported to the Australian Government in June 2021, showed that four out of five Australian adults had been exposed to misinformation about COVID-19. They also found that online misinformation, such as the propagation of anti-vaccine narratives within the Australian community, had a direct negative impact on the trust that people place in democratic institutions and public health agencies. These narratives often originate overseas but quickly spread through local communities. The fact-checking organizations that have traditionally verified statements made by public figures or politicians in public and mainstream media now must also monitor and debunk dramatically faster-spreading claims on social media platforms. Narratives containing misinformation are having a direct and negative impact on how people consume information: They may influence the content we engage with and the search terms we enter.[10] Given that an informed citizenry is a cornerstone of democracy, public decision making is at risk.

The significance of the problem was also recognized in the International Cyber and Critical Technology Engagement Strategy released by the Australian Government, which identifies digital misinformation as a clear risk to the security and safety of Australia, the Indo-Pacific region, and beyond. Countries across East Asia and Oceania introduced legislation that specifically targets so-called 'fake news' and they created voluntary codes of practice developed in partnership with the technology industry.

Despite such efforts, as of December 2022, out of the 122 currently verified signatories of the Poynter's International Fact-Checking Network (IFCN), only eight

IMAGE BY HURCA

Computational methods to assist in fact-checking.



are in the East Asia and Oceania region: Australian Associated Press (AAP) and RMIT FactLab in Australia; Cek Fakta Liputan 6, MAFINDO, Tempo.co, and Tirto ID in Indonesia; and Rappler and Verafiles Incorporated in The Philippines.[a]

Some platforms have turned to fact-checkers to help identify problematic content. However, the deluge of misinformation means that the checkers cannot handle the large number of claims that need to be assessed. Algorithmic assistance may therefore be beneficial to identifying instances of misinformation.

Computational Methods to Tackle Misinformation

In the last few years, a trend has emerged of computing professionals leveraging hybrid-intelligence (human and artificial) methods to remove misinformation from online platforms.[6] The problem is more complex than identifying and removing misinformation; such content needs to be comprehensively managed throughout the stages of its lifecycle: from creation to propagation and consumption. The computer science community has already developed technologies that can help at each stage (see the accompanying figure).

A range of methods—including social network analysis, natural language processing (NLP), information retrieval (IR), knowledge graphs, machine learning (ML) and artificial intelligence (AI), foundation models (for example, neural transformers such as BERT and GPT) and deep learning, data visualization, explanations of machine-learned model output, and advances in human-computer interaction (for example, new user experiences and interaction capabilities)—can assist but not replace humans during misinformation management processes. In all these stages, a close collaboration between experts, systems, and non-experts such as crowd workers is crucial to scaling up while maintaining the quality, agency, and accountability of the process.[6]

Human-in-the-loop fact-checking in the East Asia and Oceania region is in its infancy. While non-governmental organizations such as FirstDraft News (now the Information Futures Lab) are active in the region, more research is needed to better understand how hybrid-intelligence methods can be effectively embedded into misinformation management processes without taking agency away from experts.[9] In addition to debunking misinformation, computational methods can assist in *prebunking* processes by developing effective ways to educate people about misinformation, thus enhancing digital literacy. This will provide them with the skills to identify and question unverified information online. The low number of fact-checking organizations in East Asia and Oceania makes the support of computational methods more pressing in this region. Ensuring that this algorithmic assistance is useful, though, will also require region-specific attention.

What is 'fact' internationally is often counterfactual in East Asia and Oceania, so merely importing international information will not be effective. Examples of 'facts' that do not hold true in this region include seasonal issues; for much of the region summer is December–February. While having Christmas in summer is slightly countercultural, gatherings for these holidays did result in a major southern summer COVID-19 spike in 2021. Conversely, knowing that the influenza (and COVID-19) seasons occur in June and July in this part of the world is a key part of good public health advice. The region also predominantly sits close to the equator, so public health advice about sun exposure must be tailored. Most countries

a https://ifcncodeofprinciples.poynter.org/signatories

in the region drive on the left-hand side of the road, affecting road safety advice. There are also cultural differences within the region: The weekend can fall on different days, or many countries are predominantly Muslim, meaning Eid, not Christmas, is celebrated. Understanding the importance of COVID-19 vaccines being Halal was key to public health messaging in Melbourne, Australia. Democratic conventions are also unique: Australia has one of the highest rates of democratic participation in the world, at more than 90%, a direct result of compulsory voting. Nearby New Zealand, though, also had strong democratic participation, more than 80% without compulsory voting in the most recent election. Given these regional and local differences from global norms, local fact checking is key to ensuring an informed populace. Further, prebunking strategies must sit alongside existing educational and cultural norms. This presents the two-pronged problem of scarce local expertise and the need for localized resources, to build and evaluate algorithmic tools and human-in-the-loop solutions for the region.

The Filter Bubble Myth

While there is evidence that polarization in society dramatically escalated with the introduction of broadband Internet, the cause is not well understood. Filter bubbles, formed by algorithms delivering personalized content that reinforces a particular worldview, have become an incredibly popular explanation. The conceptualization of the filter bubble was coined in a book of the same name by Eli Pariser.[7] However, the filter bubble concept may distract from the deeper epistemic causes of polarization. Pariser's book, cited thousands of times, makes the case but provides limited evidence of bubbles being formed by search engines. Empirical studies indicate a lack of such bubbles, going further to suggest that search platforms increase exposure to contrary viewpoints. Cross-disciplinary teams of computer scientists, media specialists, information scientists, industry researchers, and psychologists are working together on the issue of search personalization through novel experimentation, which has better revealed the role that search engines play in polarization.

The Australian Search Experience,[3] a project carried out by the Australian Research Council Centre of Excellence for Automated Decision-Making and Society (ADM+S), is a data-donation study where more than 1,000 people across Australia were recruited to examine whether search engines returned different kinds of results across the cohort. Participants installed a Web browser plug-in that issued periodic queries to well-known search engines using the participants' accounts. Search results were scraped by the plug-in and returned to researchers for examination. The queries were drawn from a pre-defined list of common searches on a range of topics spanning political, controversial, and everyday categories. While the research is ongoing, initial findings indicate that, although search results were found to be contextualized to specific

Narratives containing misinformation are having a direct and negative impact on how people consume information.

Sample of query variants crowdworkers generated, drawn from the UQV test collection.[2]

Description of Information Need	Example Query Variants
A group of local farmers has been protesting outside your energy utility's offices, complaining about a plan to build a wind farm on hills near their properties. Their placards say that the disadvantages (cons) outweighed the advantages (pros). Until now, you had always assumed that wind power was a good thing; now you are not so sure and decide to find out more.	▸ Advantages and disadvantages of wind power ▸ Cons of wind power ▸ Is wind power good? ▸ Negative effects of using wind power ▸ Pros and cons of wind power ▸ Information about wind power ▸ Engineering principles of wind turbines

A trend has emerged of computing professionals leveraging hybrid-intelligence methods to remove misinformation from online platforms.

geographic locales, algorithmic personalization in search engines may be less extensive than was suggested by previous filter-bubble research. This leads to the question: If search is largely homogeneous, where is information polarization coming from?

One possible answer lies in work of our region's IR community examining the impact of query variation in search. While search-engine users have been studied for decades, recent experiments where a large number of people are asked what query they would use when seeking to satisfy a common information need have found an astonishing range of distinct queries.[2] To illustrate, the accompanying table lists a sample drawn from more than 50 query variants found when 100 crowdworkers were asked how they would search for information about wind power. Such extensive variations were recorded across a diverse set of 100 topics. The results of the experiment are packaged in a test collection that captures this user query variability (UQV).

Query variations were found to have a significant impact on search-engine performance. Wide variations in the queries submitted to commercial search engines were identified,[1] and detailed statistical analysis found that variations in queries had a substantially larger effect on search results than any change in the workings of a search algorithm.[5] The Australian Associated Press recently debunked a social media post with a false claim about the lifespan of wind farm generators.[b] One can see in the sample queries shown in the table that different queries appear to reflect different attitudes on the topic. It is natural to wonder whether misinformation influences the way people choose the keywords that they type into a search engine.

Opportunities to Address Polarization in Search Engines

This collection of findings suggests that polarization in search is being driven not by algorithms but by searchers.[1] The research trend highlights a critical oversight in search-engine algorithm design: understanding how search algorithms react to and potentially alleviate this user variation. The research challenges of such work include:

- Understanding the reasons for the variation people show when searching—for example, demographics, search habits, domain knowledge, cognitive biases, and how people are prompted to search.
- Exploring if and how people are influenced by others to search in particular ways.
- Determining how search algorithms can be adapted to better handle the variation.

Initial results suggest that the way people construct queries is informed by established searching habits, although other factors, such as existing knowledge, biases, prompts, and so on, most likely also contribute. Questions about how people are influenced to search must be examined. Here, misinformation seems to play a crucial role, and collaborations with fact-checking organizations in the region[4,8] are helping to better understand how people formulate their queries when they encounter misinformation and interventions—for example, verified content produced by fact-checkers. When examining search algorithms, the roles and responsibilities of search engines will be questioned. Most would agree that search engines should return the most reliable content, but should they intervene in trying to change user views arising from polarizing queries? Answers to such questions must be approached in a nuanced way because confronting people with views too distant to their own could alienate them. Other key research issues include investigating whether queries resulting from a disinformation campaign can be reliably detected, whether search engines could and should detect the sole pursuit of confirmatory information and evidence of confirmation biases,

b http://bit.ly/turbineFactCheck

and whether search results can be tailored to support reflection and understanding of those with beliefs different to one's own.

The Road Ahead

Detailing these two case studies shows that a richer engagement between humans and machines ensures more effective outcomes in the management of information and a better understanding of how online information is sought. The work described here represents an important emerging trend in our region, but the impact is felt far beyond this geographic area. This work also presents several grand challenges in deploying these assistive technologies at a massive scale and realizing human-AI cooperation in practice.[c]

Computing professionals must continue collaborating with other disciplines to make and integrate advances in critical areas, such as fairness, accountability, transparency, explainability, and the safety of human-AI cooperation. Misinformation and its exposure will only grow in the coming years, as will the adversarial uses of computational methods to generate and spread disinformation narratives. Polarization will persist as long as we fail to understand the causes of query variation in search-engine engagement and fail to develop more robust search algorithms capable of handling that variation. As a community, we must meet these challenges head on. Understanding and supporting the interplay of humans and algorithmic systems will ultimately lead to better outcomes for all.

Acknowledgments. The authors thank Axel Bruns, Luke Gallagher, Timothy Graham, James Meese, Stefano Mizzaro, Quoc Viet Hung Nguyen, Abdul Karim Obeid, and Falk Scholer for their contributions and feedback toward this work. This research is partially supported by the Australian Research Council (DE200100064, CE200100005, IC200100022). The authors acknowledge the Traditional Custodians of Country throughout Australia and their connections to land, sea, and community. We pay our respect to their Ancestors and Elders past and present and extend that respect to all Aboriginal and Torres Strait Islander peoples today.

References

1. Alaofi, M. et al. Where do queries come from? In *Proceedings of the 45th Intern. ACM SIGIR Conf. Research and Development in Information Retrieval* (2022), 2850–2862; https://doi.org/10.1145/3477495.3531711.
2. Bailey, P., Moffat, A., Scholer, F., and Thomas, P. UQV100: A test collection with query variability. In *Proceedings of the 39th Intern. ACM SIGIR Conf. Research and Development in Info Retrieval* (2016), 725–728; https://doi.org/10.1145/2911451.2914671.
3. Bruns, A. *Australian Search Experience Project: Background Paper.* Technical Report. ARC Centre of Excellence for Automated Decision-Making and Society (2022); https://doi.org/10.25916/k7py-t320.
4. Cerone, A. et al. Watch 'n' Check: Towards a social media monitoring tool to assist fact-checking experts. In *Proceedings of the 2020 IEEE 7th Intern. Conf. Data Science and Advanced Analytics*, 607–613; https://doi.org/10.1109/DSAA49011.2020.00085.
5. Culpepper, J.S., Faggioli, G., Ferro, N., and Kurland, O. Topic difficulty: Collection and query formulation effects. *ACM Trans. Inf. Syst. 40*, 1, Article 19 (Sept. 2021); https://doi.org/10.1145/3470563.
6. Demartini, G., Mizzaro, S., and Spina, D. Human-in-the-loop artificial intelligence for fighting online misinformation: Challenges and opportunities. *IEEE Data Eng. Bull. 43*, 3 (2020), 65–74; http://sites.computer.org/debull/A20sept/p65.pdf.
7. Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin, U.K. (2011).
8. Saling, L.L. et al. No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. *PLOS ONE 16*, 8 (August 2021), 1–13; https://doi.org/10.1371/journal.pone.0255702.
9. Thomson, T.J. et al. Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. *Journalism Practice 16*, 5 (2022), 938–962; https://doi.org/10.1080/17512786.2020.1832139.
10. Yom-Tov, E., Dumais, S. and Guo, Q. Promoting civil discourse through search engine diversity. *Social Science Computer Review 32*, 2 (2014), 145–154; https://doi.org/10.1177/0894439313506838

Damiano Spina is a senior lecturer of data science and DECRA Fellow at RMIT University, Melbourne, Australia.

Mark Sanderson is a professor of information retrieval and Dean of Research & Innovation at RMIT University, Melbourne, Australia.

Daniel Angus is a professor of digital communication at Queensland University of Technology, Brisbane, Australia.

Gianluca Demartini is an associate professor of data science at The University of Queensland, Brisbane, Australia.

Dana McKay is a senior lecturer of data science at RMIT University, Melbourne, Australia.

Lauren L. Saling is a senior lecturer of psychology at RMIT University, Melbourne, Australia.

Ryen W. White is a general manager and deputy lab director at Microsoft Research, Redmond, WA, USA.

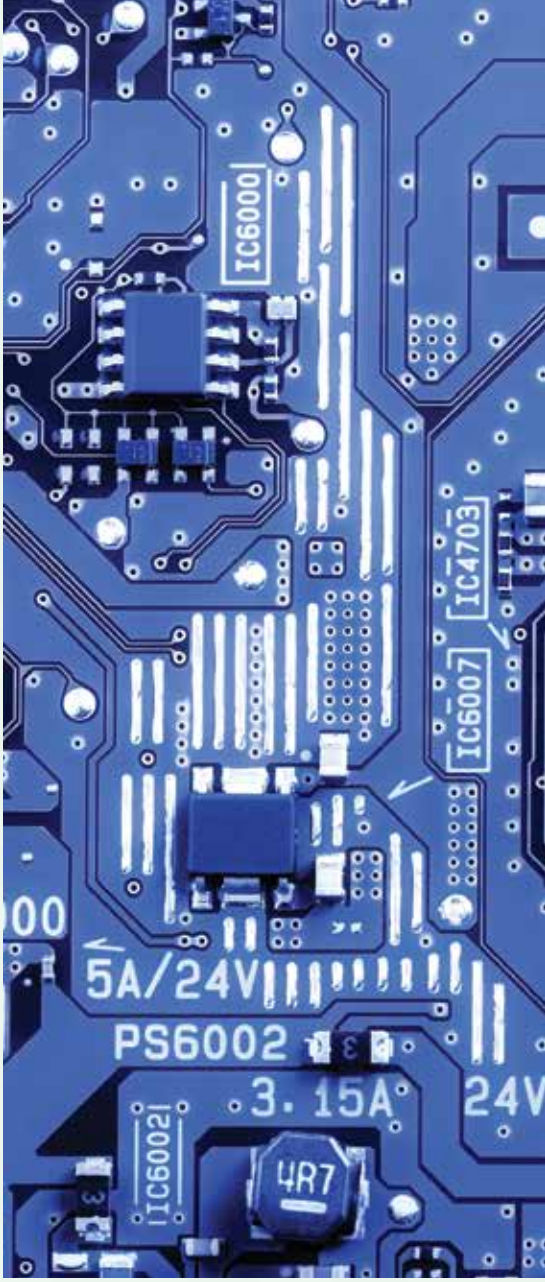
Computing professionals must continue collaborating with other disciplines to make and integrate advances in critical areas, such as fairness, accountability, transparency, explainability, and the safety of human-AI cooperation.

c Also highlighted in the ACM Technology Policy Council's Statement on Principles for Responsible Algorithmic Systems; https://bit.ly/jointAIstatement

DOI:10.1145/3587264

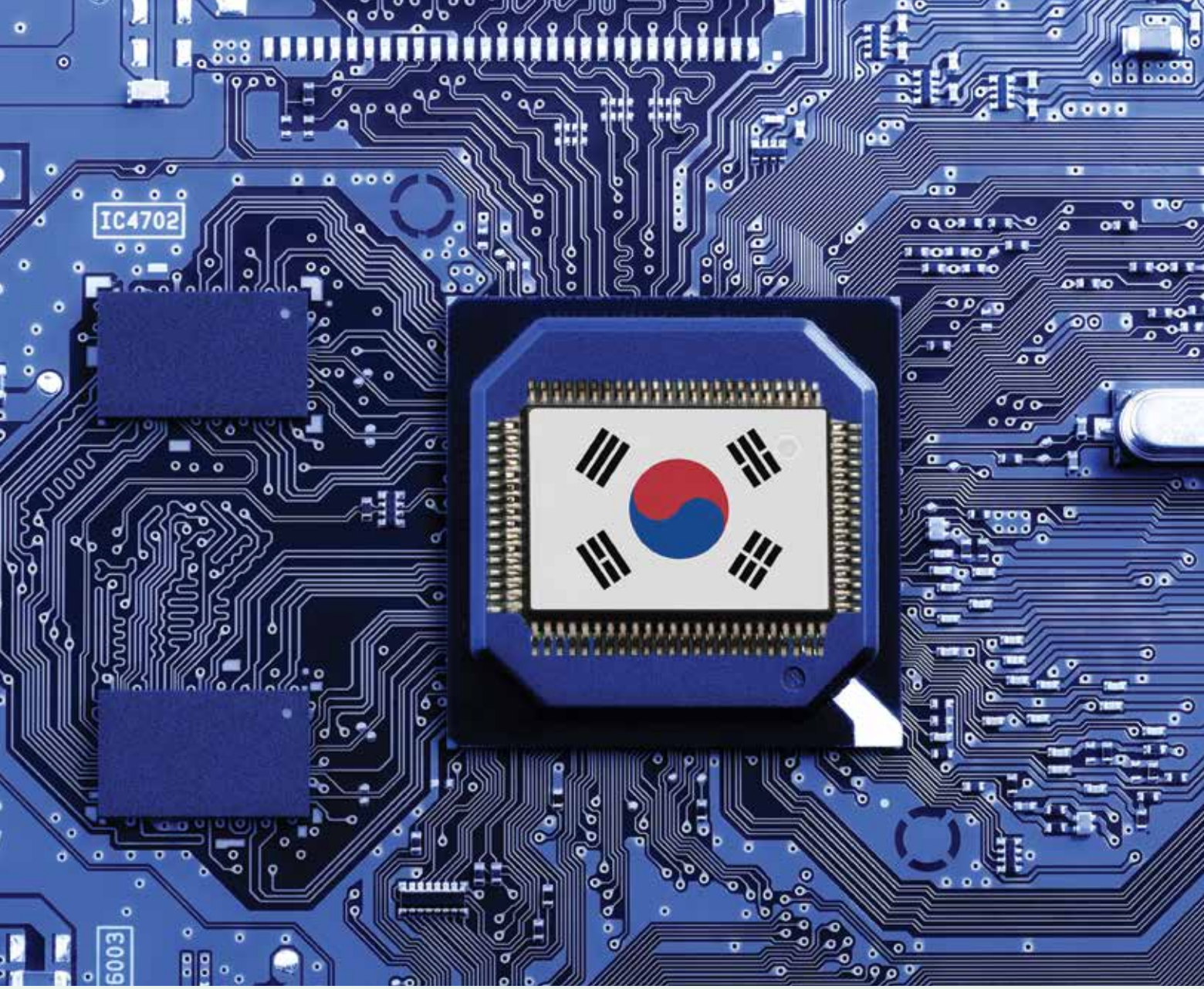
BY JI-HOON KIM, SUNGYEOB YOO, AND JOO-YOUNG KIM

South Korea's Nationwide Effort for AI Semiconductor Industry



AS GLOBAL COMPETITION in the semiconductor industry has intensified with trade conflicts and semiconductor shortages, major countries worldwide have started to work on their government policy and investment plan to win technological hegemony. South Korea's semiconductor industry, which makes up almost 20% of the nation's gross domestic product (GDP), is heavily concentrated on the memory semiconductor sector.[5] It dominates the global memory semiconductor market with a 56.9% share but has little influence on the other sectors of the industry, including logic, analog, and optical discrete, where it has less than a 3% market share. To grow the nation's biggest industry further, South Korea has put a priority on non-memory sectors. The emerging AI chip market is an especially great opportunity for them, as demand is expected to increase exponentially with the paradigm shift in computing. South Korea aims to become a comprehensive semiconductor powerhouse from this colossal opportunity. In this article, we report recent nationwide efforts taking place in South Korea to challenge the emerging AI semiconductor industry. We organize these efforts in four sections (government, major companies, fabless startups, and academia), while Figure 1 overviews the overall efforts.

The South Korean government's K-semiconductor strategy is to build the world's best semiconductor supply chain by 2030 with a $450B investment plan.[2] Learning from the semiconductor crisis, the goal is to stabilize and internalize the semiconductor supply chain by gathering fabless, foundry, and packaging companies in a clustered area called



IMAGERY FROM SHUTTERSTOCK.COM

the K-Semiconductor Belt for seamless silicon product manufacturing. The government has also announced nearly $260B-worth of tax deductions for semiconductor facilities and R&D funding,[3] in addition to making the approval process for the expansion of semiconductor manufacturing facilities faster and more flexible, and subsidizing up to 50% of the construction cost of the facility's power infrastructure. The government pledged to invest in strategically important semiconductor sectors, including power, automotive, and AI semiconductors, as part of its long-term R&D roadmap. With the above financial and regulatory support, the government set a goal of doubling semiconductor production to $245B, with an export target of $200B, by 2030.

The government also is trying to build a collaborative ecosystem that helps once-segmented academia and industry work together to make competitive products for the global market. It has set up national programs to facilitate collaboration between university labs and startup companies by funding technology transfer and commercialization. The government also is subsidizing startup fabless companies to use the latest EDA tools and semiconductor process technologies.

Lastly, the government is heavily focused on fostering high-quality manpower for the semiconductor industry. Based on a survey that found at least 270,000 persons will be needed to sustain the industry in 10 years, the government aims to foster 150,000 more semiconductor experts at all levels (junior colleges, undergraduate schools, and graduate schools) by 2030. It also plans to establish new research centers and university departments to develop competent researchers and experts in the semiconductor field.

Major countries, including the U.S. and China, also have announced support policies for their semiconductor industries. The U.S. passed the CHIPS Act in 2022, which promises 25% tax credits and a $52B investment to enhance domestic semiconductor manufacturing. China's "Made in China 2025" initiative announced in 2015 has aimed to lift the country's chip production from less than 10% of demand to 70% in 2025, fostering government-backed fabless and manufacturing companies. In contrast, South Korea's plan is more

Figure 1. Overview of South Korea's nationwide efforts for the AI semiconductor industry.

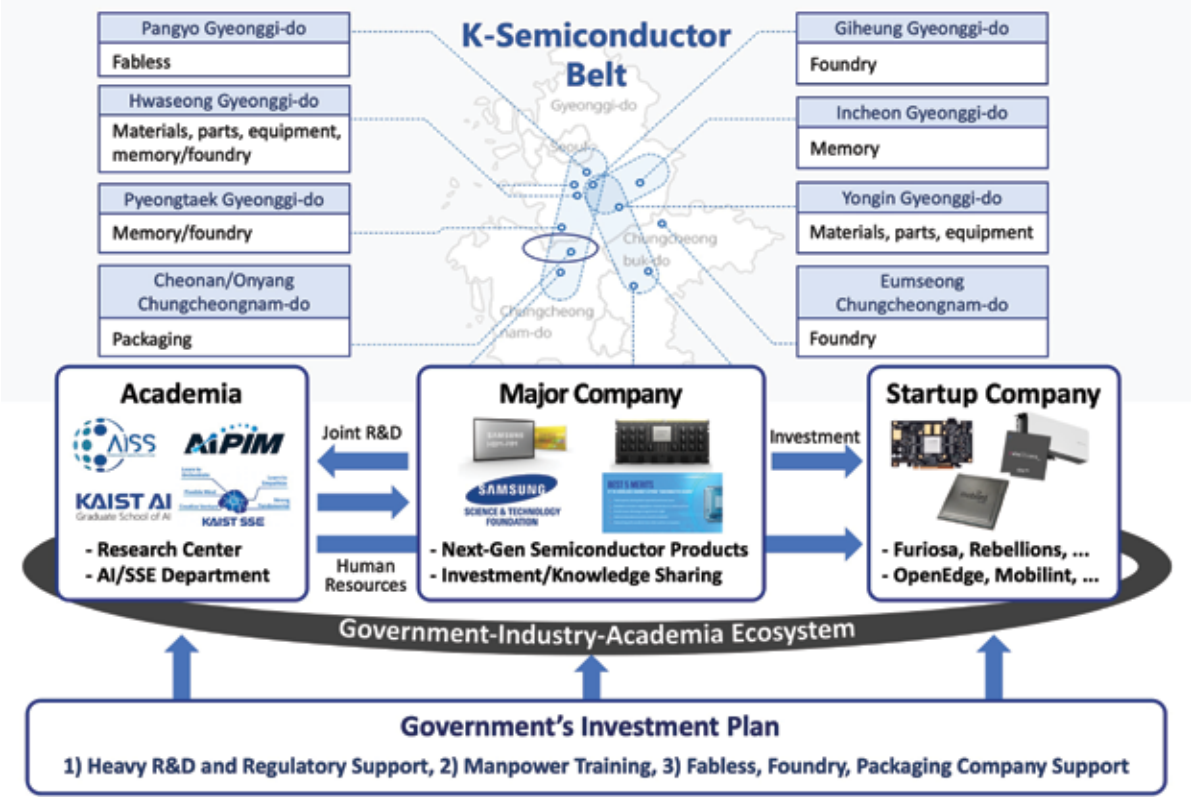
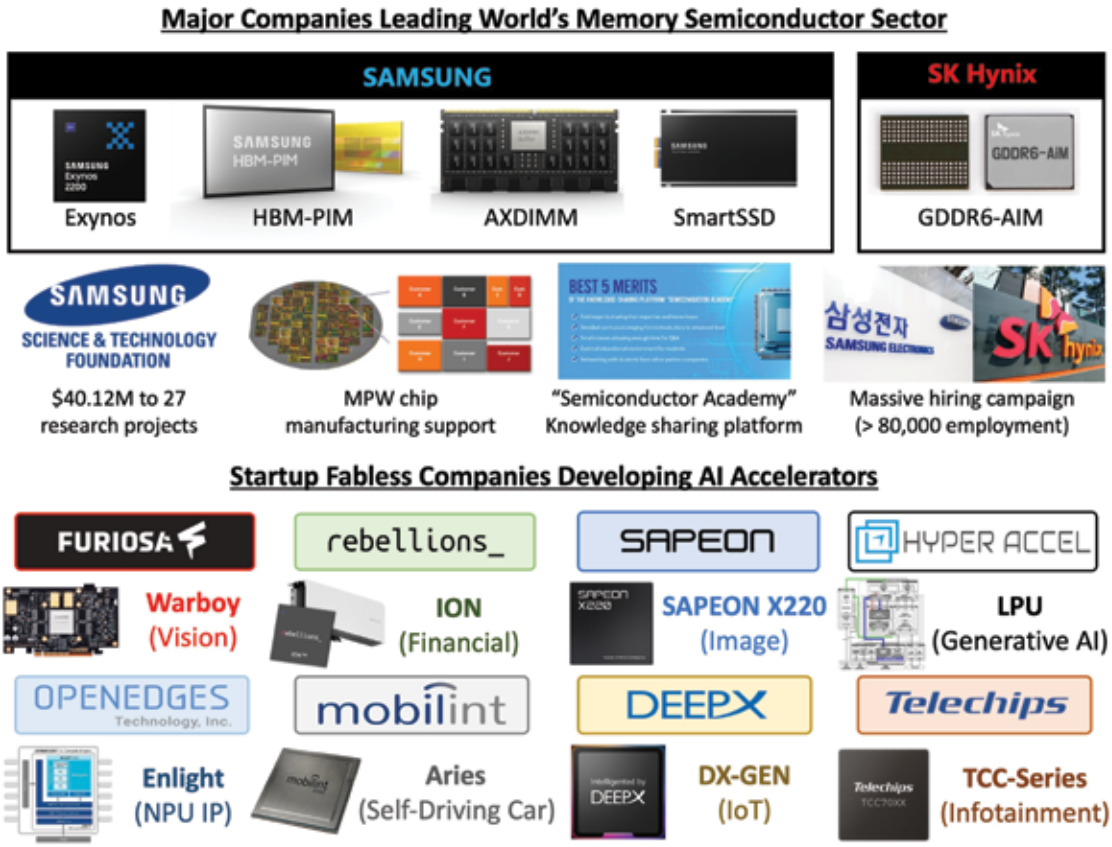


Figure 2. Summary of South Korea's AI semiconductor products.



concentrated and streamlined, emphasizing human resource development overall.

Major Companies

South Korea's chip companies working in AI semiconductors can be divided into two groups: major companies leading the world's memory semiconductor market, and newly founded fabless startup companies developing AI accelerators. Figure 2 summarizes their major products and activities related to AI semiconductor development.

Samsung Electronics and SK Hynix, the world's leading memory semiconductor companies, have launched investment and employment plans for their AI semiconductor and foundry businesses. Samsung Electronics is trying to develop next-generation AI semiconductor products by leveraging its strengths in mobile chipset design and memory manufacturing. Samsung develops its own neural processing units (NPUs) and integrates them into multiple processing platforms, including the Exynos mobile processor and the Exynos auto processor series.

Another notable product development direction is putting AI computation logic into memories. Both Samsung and SK Hynix are developing intelligent memory chips using in/near-memory processing technology, which can mitigate the system bandwidth bottleneck of conventional von Neumann architecture. Samsung released the HBM-PIM, which incorporates the AI processing function into the HBM2 memory stack, and the AXDIMM, which adds an AI engine inside the buffer chip of the traditional DIMM form factor. The company also developed SmartSSD, which integrates Samsung's solid-state drive (SSD) and Xilinx's field-programmable gate array (FPGA) chip on a single card, with a fast direct data path between them for near-data processing. SK Hynix recently released the GDDR6-Accelerator-in-Memory (AiM) chip that integrates processing units in the latest GDDR6, which can provide a total of 1TFLOPS processing throughput.

In addition, once secretly focusing solely on their own product development, major companies are starting to open relations with the academia and research communities to learn the latest AI technologies and engage with well-educated researchers. For example, the Samsung Science & Technology Foundation announced grants totaling $40.12 million to 27 high-profile research projects, bolstering collaboration with academia.[4] SK Hynix also launched a knowledge-sharing platform named "Semiconductor Academy," which provides online/offline lectures to university students, job seekers, and employees, covering a diverse range of topics including semiconductor basics, devices, and design. Moreover, both Samsung and SK Hynix continuously provide multi-project wafer (MPW) services for startup fabless, design houses, and academia. This allows greater foundry selection for domestic fabless companies and enables competitive prototyping of their AI chips with the latest process technologies. Samsung has announced a massive hiring plan to employ more than 80,000 people in new jobs, as the demand for AI semiconductor products continues to increase.

Fabless Startups

Ever since it launched the Ministry of SMEs and Startups in 2017, South Korea has been investing heavily in building a startup ecosystem to pivot from a conglomerate-based to a startup-based industry. With successful supporting programs such as K-Unicorn and Tech Incubator Program for Startups (TIPS), South Korea has become a good place for startups, surpassing $6.4B in venture capital funding in 2021.[6] Benefiting from this environmental change, more than 10 fabless startups targeting AI hardware accelerators have been founded in South Korea in the past five years. Unlike the major companies that develop next-generation AI products based on existing products, the fabless startups develop AI accelerators from scratch for specific application domains. They can be segmented into two groups: the companies targeting datacenter applications (FuriosaAI, Rebellions, Sapeon, and HyperAccel) and companies targeting edge applications (OpenEdge, Mobilint, DeepX, and Telechips).

FuriosaAI is one of the first fabless startups to develop AI accelerators for high-performance vision tasks such

Once secretly focusing solely on their own product development, major companies are starting to open up relations with the academia and research communities to learn the latest AI technologies and engage with well-educated researchers.

Table 1. Industry-contracted semiconductor departments in universities.

University	Contracting Company	# of Students	Year of Establishment
Sungkyunkwan University	Samsung Electronics	70	2016
Yonsei University	Samsung Electronics	50	2019
Korea University	SK Hynix	30	2021
Sogang University	SK Hynix	30	2022
Hanyang University	SK Hynix	40	2022
POSTECH	Samsung Electronics	40	2022
KAIST	Samsung Electronics	100	2022

Table 2. Technical challenges.

AI Accelerators	Processing-in-Memory (PIM) Semiconductors
High-performance Tensor Hardware Design, Full-stack Software Development, Latest Logic Process (< 14nm)	In-Memory Logic/Circuit Design, System Software for PIM Hardware, Memory Process (Especially DRAM)

as image classification and object detection in datacenters. The company released the Warboy prototype in 2021, proving it could achieve 1.5x higher performance and 4x higher performance per price than a comparable GPU, NVIDIA's T4, for target applications.

Rebellions is a fabless startup company developing an AI accelerator for high-frequency trading (HFT). That company released its first accelerator card for intelligent HFT, called LightTrader, in 2022; it achieves 64 trillion operations per second (TOPS)/16 trillion floating-point calculations per second (TFLOPS) peak performance, with workload scheduling and dynamic voltage-frequency scaling.

Sapeon also is developing its own AI inference chip for low-latency AI inference on image data. In 2022, Sapeon released the X220, the first commercialized AI semiconductor chip, showing 2.3x and 2.2x higher performance and power efficiency than NVIDIA's A2, respectively, in the MLPerf datacenter inference benchmark.

HyperAccel is the newest of the startups, founded this year to target the acceleration of hyperscale AI models such as OpenAI's GPT. HyperAccel uses multiple AI accelerators with memory-bandwidth maximization for emerging generative AI workloads.

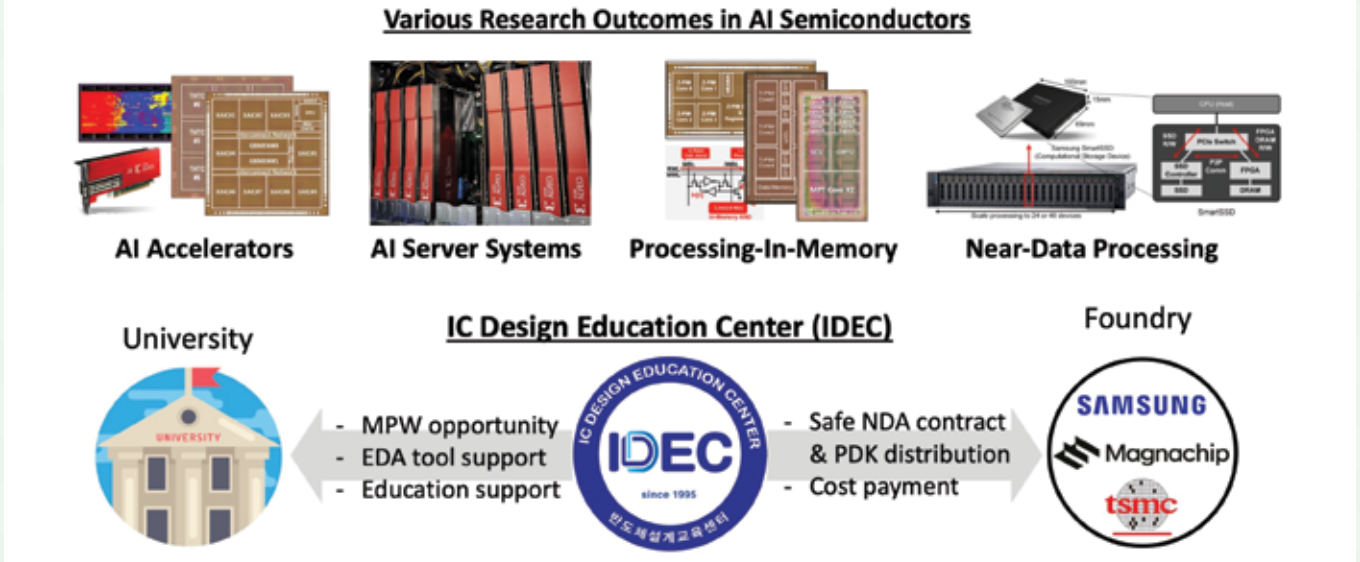
OpenEdges Technology is developing low-power, high-performance NPU IP and related high-speed memory systems. The company's strength in memory controllers and subsystems makes them promising for NPU development. Last year, OpenEdges released Enlight, an AI accelerator IP with an optimized network model compiler that minimizes DRAM traffic.

Established in 2019, Mobilint is an AI semiconductor startup developing high-performance AI inference NPUs for edge devices and cars. Its NPU implementation on FPGA took first place in the global benchmark MLPerf for two consecutive years (2020 and 2021). Mobilint last year unveiled its Aries chip, which delivers up to 80 TOPS of AI performance at maximum capacity.

DeepX is a startup company developing power-efficient NPU chips for Internet of Things (IoT) devices. The company's DX-L1 system on a chip targets relatively simple IoT applications, such as smart camera sensors and security cameras, while the DX-H1 targets huge IoT applications, such as smart factories and smart buildings.

Lastly, Telechips is a medium-sized company developing intelligent automotive solutions for autonomous vehicles. The company has released various infotainment application processors for audio, digital cluster, and cockpit systems with low power requirements

Figure 3. Research outcomes of South Korean universities and IDEC's educational support.



and a high level of security.

South Korean fabless startups aim beyond the domestic market to enter overseas markets with their AI accelerators, leveraging the country's competitive IT infrastructure as a testbed. They are targeting developed markets such as the U.S. and Europe, but also are looking at lucrative emerging markets, such as India and Southeast Asia.

Universities and Education

Korean universities are making great efforts to expand their research on AI chips and on next-generation semiconductor technology. Since Sungkyunkwan University established its contracted semiconductor department with the help of Samsung Electronics in 2006, Yonsei and Korea universities have implemented similarly contracted departments with Samsung in 2019, and with SK Hynix in 2021. In such industry-contracted semiconductor departments, the curriculum is focused on semiconductor engineering, including its basic devices, design, and system integration, to educate students to be ready for work in the domain at graduation. They also get hands-on project experience and mentorship opportunities from industry experts. After graduation, the students are guaranteed mployment by the contracting company, allowing many talented experts to work in the domestic industry. As summarized in Table 1, seven major universities in South Korea have industry-contracted semiconductor departments, and more than 350 students will be raised to the level of AI semiconductor experts soon.

The influence of the Korea Advanced Institute of Science and Technology (KAIST) is especially critical to South Korea's semiconductor industry, as described in the biographic article of Choong-Ki Kim.[1] KAIST founded its semiconductor system engineering department, including 50 prominent professors, in 2022, with three specialized tracks—semiconductor device/process, chip design and VLSI, and system software and algorithms—to lead and transform the industry towards AI/system-focus, rather than remaining memory-focused. The school also has attracted governmental research centers in the AI semiconductor domain, such as the AI Semiconductor System (AISS) research center and the Processing-in-Memory (PIM) research center. Involving more than 10 research labs and 100 graduate students, each research center conducts innovative research over a long-term period, nurturing human resources with technical competence.

South Korean universities show top-notch research capabilities in the areas of AI semiconductors. As illustrated in Figure 3, they result in strong research outcomes in AI accelerator architecture and design, server-scale AI systems, processing-in-memory chips, and near-data processing. They publish many papers in top-tier conferences and journals, such as the International Solid-State Circuits Conference, the International Symposium on Computer Architecture, the IEEE/ACM International Symposium on Microarchitecture, and the *IEEE Journal of Solid-State Circuits*.

The IC Design Education Center (IDEC) is the unsung hero of South Korea's semiconductor industry, helping universities pursue research in the chip design field, which is expensive due to manufacturing costs. Since 1995, IDEC has supported MPW programs and access to essential electronic design automation (EDA) tools at low cost, as well as providing practical education to universities. It manages sensitive non-disclosure agreements (NDAs) and process design kit (PDK) distribution between universities and foundry companies such as Samsung, TSMC, and Magnachip. More specifically, they serve more than 10 MPW shuttles and 50 EDA tools from over 15 vendors to universities every year. In addition, IDEC annually holds a chip design contest to exhibit silicon chips produced through its MPW programs to the community and industry, sharing the latest research outcomes for enhancing domestic chip design competitiveness. It has been operating more than 50 online/offline lectures on chip design every year, and recently started an accelerated degree program for industry practitioners.

Conclusion

South Korea is striving to become a comprehensive semiconductor powerhouse by preempting the emerging AI semiconductor market via nationwide efforts among government, industry, and academia. Table 2 summarizes technical challenges for their two target products: AI accelerators and PIM semiconductors. For AI accelerators, efficient software and hardware execution through full-stack development and high fabrication costs are the challenges. For PIM semiconductors that integrate logic inside the memory, low-level circuit innovation and system software to enable their adaptation in the system are important, while their fabrication is primarily possible only by memory vendors.

To this end, the government has a bold investment plan with both financial and regulatory support, promoting the development of a collaborative ecosystem by 2030 in which fabless, foundry, and packaging companies, as well as universities, will work together to enable seamless silicon product manufacturing. The government also plans to aggressively foster AI semiconductor experts by initiating school departments and research centers. Memory chip giants such as Samsung and SK Hynix are diversifying their product lines with their own NPU IPs and in/near-memory processing technology. More than 10 fabless startup companies have been founded recently and are actively developing AI acceleration solutions for various application domains. C

References

1. Kim, D-W. The Godfather of South Korea's chip industry: Kim Choong-Ki's "Engineer's Mind" helped make the country a semiconductor superpower. *IEEE Spectrum 59*, 10 (2022), 32–38.
2. Kim, J. South Korea plans to invest $450bn to become chip 'powerhouse'. *NIKKEI Asia*, http://bit.ly/3Yvbm02.
3. Lee, J-Y. S. Korea targets W340tr investment for chip supremacy. *The Korea Herald*; https://www.koreaherald.com/view.php?ud=20220721000751.
4. Lee, S. Samsung announces research projects for $40.12 million sponsorship. *Pulse News*; https://pulsenews.co.kr/view.php?year=2022&no=305553.
5. Rousselot, S. The ambiguous position of the South Korean semiconductor industry in the US-China tech war. *Asia Power Watch*; http://bit.ly/3mrFeNz.
6. Yoon, S. This is how South Korea can become a global innovation hub. *The World Economic Forum*; http://bit.ly/3ZuRSKm.

Ji-Hoon Kim is a Ph.D. student in the School of Electrical Engineering at Korea Advanced Institute of Science & Technology (KAIST), Daejeon, South Korea.

Sungyeob Yoo is a Ph.D. student in the School of Electrical Engineering at Korea Advanced Institute of Science & Technology (KAIST), Daejeon, South Korea.

Joo-Young Kim is a professor in the School of Electrical Engineering at Korea Advanced Institute of Science & Technology (KAIST), Daejeon, South Korea.

DOI:10.1145/3588591

BY TAO LUO, WENG-FAI WONG, RICK SIOW MONG GOH, ANH TUAN DO, ZHIXIAN CHEN, HAIZHOU LI, WENYU JIANG, AND WEIYUN YAU

Achieving Green AI with Energy-Efficient Deep Learning Using Neuromorphic Computing

DEEP LEARNING (DL) systems have been widely adopted in many industrial and business applications, dramatically improving human productivity, and enabling new industries. However, deep learning has a carbon emission problem.[a] For example, training a single DL model can consume as much as 656,347 kilowatt-hours of energy and generate up to 626,155 pounds of CO_2 emissions, approximately equal to the total lifetime carbon footprint of five cars. Therefore, in pursuit of sustainability, the computational and carbon costs of DL have to be reduced.

Modeled after systems in the human brain and nervous system, neuromorphic computing has the potential to be the implementation of choice for low-power DL systems. Neuromorphic computing features both neuromorphic algorithms, called spiking neural networks (SNNs), and neuromorphic hardware which are dedicated ASICs optimized for SNNs. Spiking neural networks are regarded as the third generation of artificial neural networks (ANNs), in which spikes (represented by "0" and "1" in the computing system, where "0" means the absence of a spike) are used to transmit information between neurons. With such a spiking mechanism, costly multiplications could be replaced by more energy-efficient additions, mitigating the intensity of the computation. Neuromorphic hardware, on the other hand, has a non-von Neumann "processing in memory" architecture, where computations are integrated into or near a distributed memory architecture. Combined with promising emerging memory devices such as non-volatile resistive and magneto-resistive memories (that is, RRAM and MRAM) to store synaptic weights, both static power and power consumed by data movement are significantly reduced.

Though neuromorphic computing is still in its infancy, the market is expected to grow from ~$200M in 2025 to ~$20B in 2035.[b] Significant efforts have been devoted to neuromorphic computing research. As shown in Figure 1(a), players in both academia, including Stanford, and industry, including Intel and IBM, have developed neuromorphic computing systems.

To promote research in neuromorphic computing, Singapore launched an ambitious program in 2017 covering everything from

a http://bit.ly/3YzaDet

b http://bit.ly/3mDqviU

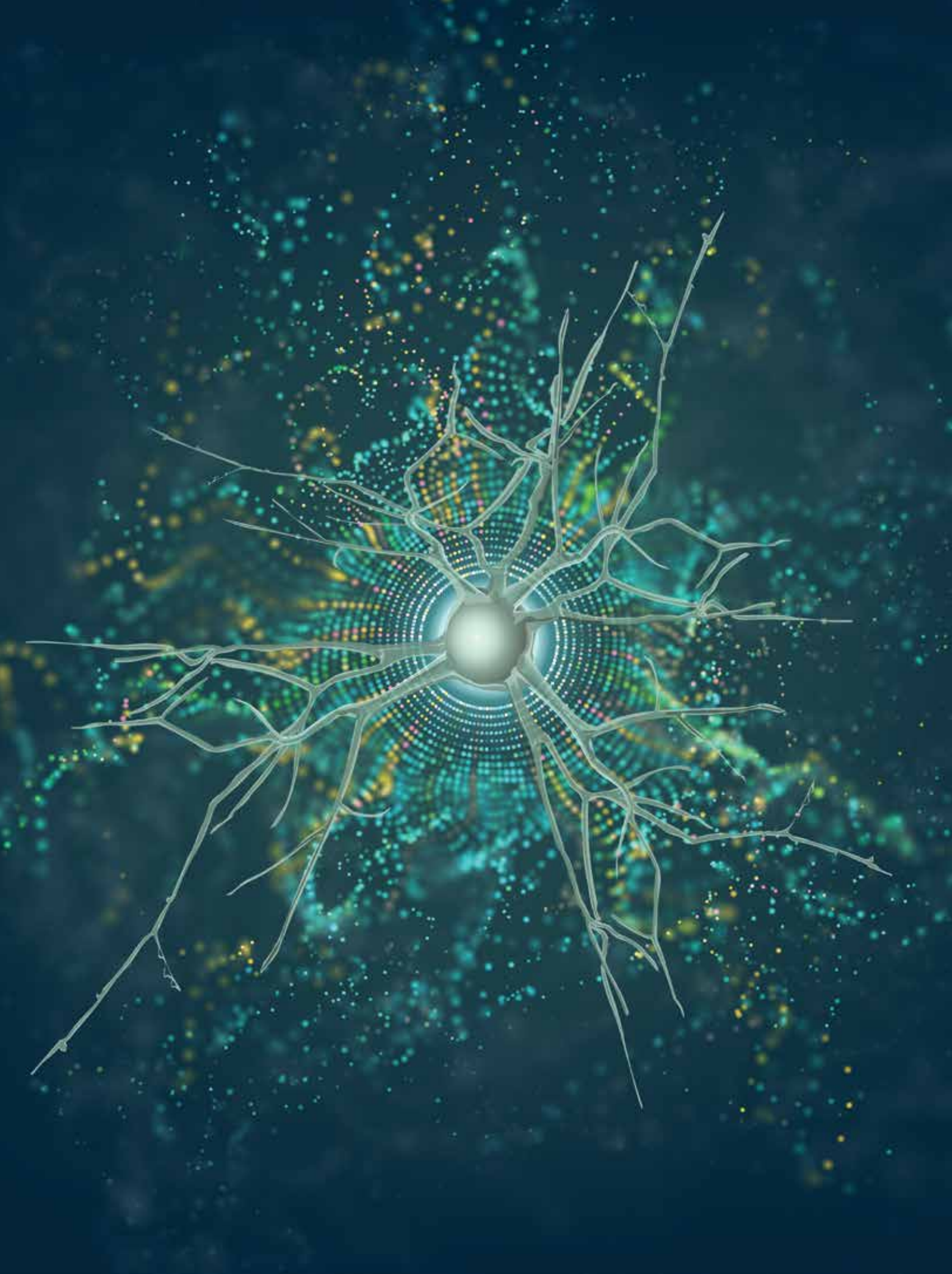
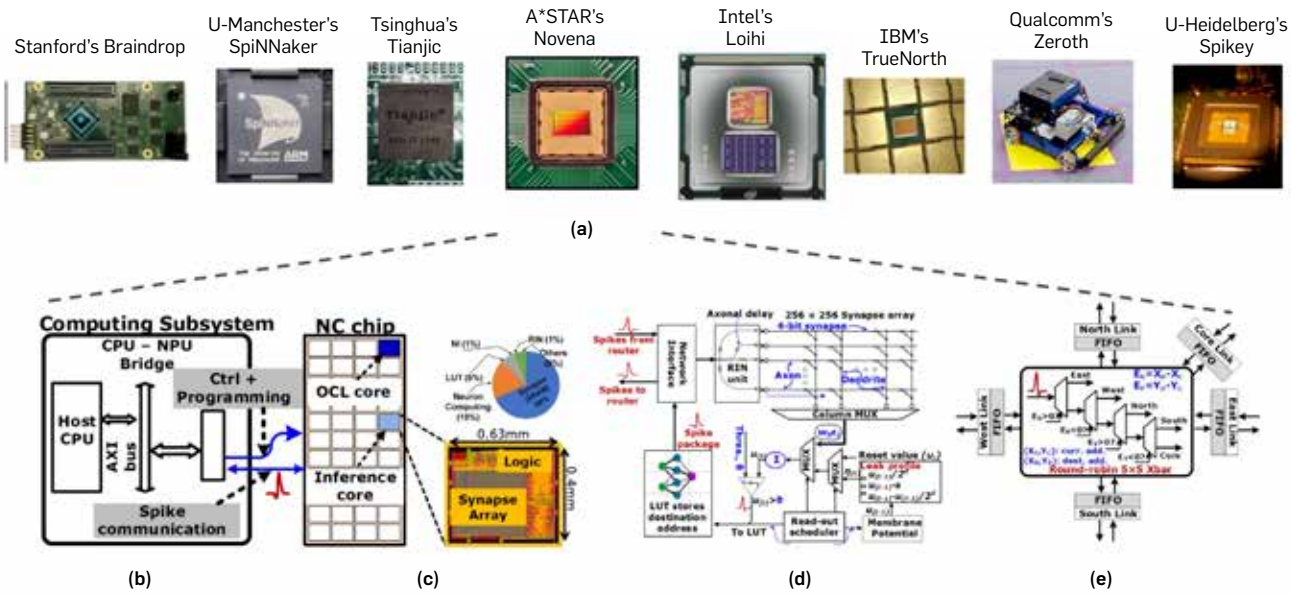


Figure 1. Neuromorphic Computing Projects Zoo.

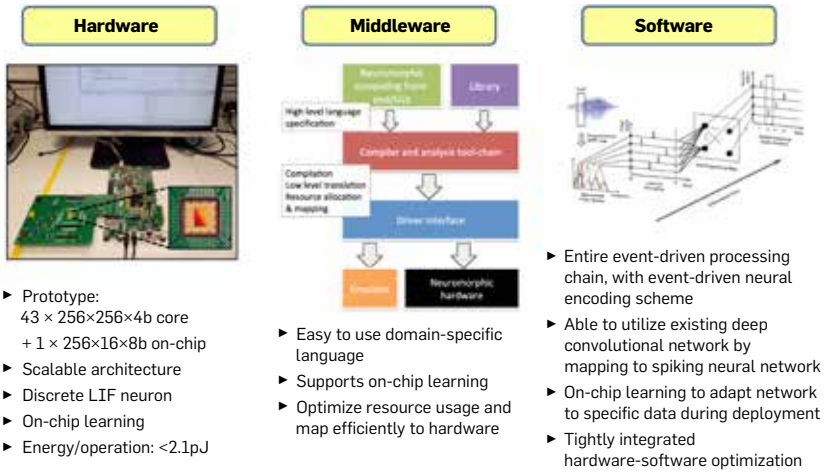
Includes (a) A*STAR's Novena, (b) Overall Novena chips organization and the host CPU Core layout, (c) Area breakdown of each core, (d) Details microarchitecture of each core design, (e) On-chip router supporting 5-input-5-output communication.



hardware to middleware to software, as shown in Figure 2. Through collaborations between the Agency for Science, Technology and Research (A*STAR) and the National University of Singapore (NUS), this program developed an end-to-end neuromorphic computing solution with < 2.1 pJ energy per synaptic operation achieved on our ASIC neuromorphic computing (NC) chip code-named Novena, advancing the current neuromorphic computing research within Singapore.

Designed to develop application-driven solutions for real-world problems, the vision of this program is for every conventional von Neumann computer in the near future to be augmented with a neuromorphic co-processor to handle big data such as text, speech, images, video, and bio-signals that require DL-related solutions. Singapore, as a popular regional hub for datacenters, is working to develop a smart nation while ensuring sustainability as part of its Research, Innovation and Enterprise 2025 Plan (RIE2025). The sheer number of smart devices running DL-related algorithms in datacenters or at edge terminals is expected to have a significant impact on the total carbon footprint of computing. Reducing the energy consumption of such devices is therefore imperative to attain green AI. The project is Singapore's contribution to the advancement of sustainable science and engineering that stands to benefit all of humanity. In this article, we introduce this program and its four aspects, including hardware, middleware, software, and system integration.

Figure 2. Features of the hardware, middleware, and software for the neuromorphic computing program.



Hardware

The Novena chip was fabricated in a 40-nm CMOS process, occupying a total area of 3.6×5.4 mm^2. Figure 1(b) shows the overall Novena chip[4,5] organization and a host CPU. The CPU configures the NC chip via a separate programming interface. A separate 64-bit bus is used for spike communication. The chip has

44 cores, 43 of which are inference cores while the last is the on-chip learning (OCL) core powered with circuit implementation of an OCL algorithm called Delta Spike Time Dependent Plasticity (Delta STDP). Delta STDP is fundamentally an error-modulated supervised STDP that is customized to perform few-shot learning only on a single layer (the last output layer) of an SNN, where the error is the deviation from target output spike count. Details of the Delta STDP circuits and algorithm can be found in Wong et al.[5] Each inference core has 256 neurons, 256×256 synapses, and 4-bit neurosynaptic weight while the OCL core has 16 neurons, 256×16 synapses, and 8-bits neurosynaptic weight. Details of the area breakdown of each inference core is shown in Figure 1(c).

Figure 1(d) details the neuronal circuit which consists of a 256×256 synaptic crossbar, neuron computation circuit, a look-up table (LUT), and a network interface. Ultra-high-V_{th} is used in SRAM cells, reducing more than 60% of total chip leakage power. To save area, neuron computations such as leak, integrate, and fire are time-multiplexed by a single circuit. To support different applications, neuron leakage profile (that is, fractional or linear) and membrane threshold value θ are configurable. We constrained our fractional leakage to $1-(1/2^d)$ so that a bit-shift operator can be used instead of a full division logic to further reduce power and area. Once the membrane potential u_i of a neuron exceeds its threshold θ, a spike will be generated, and u_i is then reset to a default level which is usually zero. By looking up its address in the LUT, a spike will be sent to the corresponding destination neuron. Output addresses from the lookup table will be queued at the network interface buffer and will only be sent to the corresponding router when it can be assured that no spikes will be dropped.

Figure 1(e) shows the block diagram of the router and neuronal circuit design. The router consists of a round-robin arbiter, interface links, XY routing algorithm logic, and crossbar switch. Each router communicates directly with its own core and sends/receives spikes to/from its four neighbors via its four ports. It is capable of handling spikes as well as debug packets, and can handle indirect addressing for partial summation configurations. The router uses handshaking protocol and thus enables globally asynchronous locally synchronous (GALS) operations for lower power consumption by removing all timing constraints and power overhead on the global clock tree at the top-level integration.

Middleware

Designing a neuromorphic processor with RRAM synaptic memory, on-chip learning circuits, and an architecture that allows system scaling for applications of increasing complexity is not easy. It requires careful software and hardware co-design, with careful considerations to be made on many design choices, with respect to the performance, energy, and area constraints. After the hardware is designed and fabricated, there is also a need for end users to easily program and make use of the chip. To tackle these challenges, in this program, we developed a middleware component to support the development of neuromorphic processors and bridge the gap between applications and the hardware platform.

Figure 2 also shows the architecture of the middleware for the neuromorphic hardware, whose components are specialized to the neuromorphic computing chip. Firstly, a system-level simulator was developed that uses CPUs and GPUs to perform neural core simulation and network-on-chip simulation.[2] It supports simulation of different RRAM material and various RRAM characteristics including stuck-at faults, random telegraph noise, and write variability. The simulator is developed to have high scalability and to support simulation of a neuromorphic chip with up to around 20,000 neural cores that was tested to run on 512 Nvidia A100 GPUs. In addition to the simulator, an FPGA-based hardware emulator for the neuromorphic chip was also developed to accelerate the simulation, which

The sheer number of smart devices running deep learning-related algorithms in datacenters or at edge terminals is expected to have a significant impact on the total carbon footprint of computing.

Surrogate gradient-based learning algorithms have been proposed to resolve the non-differentiable spike function by introducing continuous surrogate derivatives.

achieves ~2,000× speedup compared with multi-threaded execution on the CPU simulator.[3]

Secondly, to allow end users to easily program and use the chip, an end-to-end design framework for neuromorphic computing was developed. It includes a design front-end software compatible with the mainstream design framework including PyTorch and TensorFlow, with extensions to facilitate the design and training of SNNs, and a compilation middleware and analysis tool chain that compiles SNNs from the design front-end to produce configuration data required by our neuromorphic chip. It also optimizes the usage of hardware resources on the neuromorphic chip.[7]

Finally, with the simulator/emulator and end-to-end design framework, software/hardware co-exploration is performed. As the first work of its kind, we successfully tested the ImageNet dataset on a hardware-aware model on our neuromorphic chip architecture using 9,074 neural cores, demonstrating the advantages of our neuromorphic system over the state-of-the-art, achieving promising performance in terms of accuracy, number of neural cores, latency, and energy cost.

Software

Due to the non-differentiable spike function and the complex temporal dependence between spikes, how to efficiently train deep SNNs remains an open question. Various learning algorithms have been proposed. ANN-to-SNN conversion methods transform the knowledge from trained ANNs to SNN counterparts to achieve low-power and low-latency in the inference process.[6] However, ANN-to-SNN methods are unable to yield SNNs that can deal with sequence data processing. Inspired by back propagation through time, surrogate gradient-based learning algorithms have been proposed to resolve the non-differentiable spike function by introducing continuous surrogate derivatives, which however require large computing and memory resources due to frequent update of the synaptic weights at every time step. Spike-driven learning algorithms train the deep SNN in an event-driven manner, in which the spike timing is regarded as a relevant signal for synaptic weights updating.[8]

In this program, we made great progress in training deep SNNs by majorly proposing tandem learning[6] and spike timing dependent back-propagation learning (STDBP).[8] The tandem learning applies rate-based coding and transforms the ANN knowledge to the coupled SNN in a layer-wise manner to reduce conversion errors, achieving competitive performance on both frame-based and event-based datasets. The STDBP places the information in the timing of a single spike, namely temporal coding, and both the inference and learning are in an event-driven manner. STDBP achieves state-of-the-art 99.5% accuracy on the Caltech face/motorbike dataset among spike-driven learning algorithms.

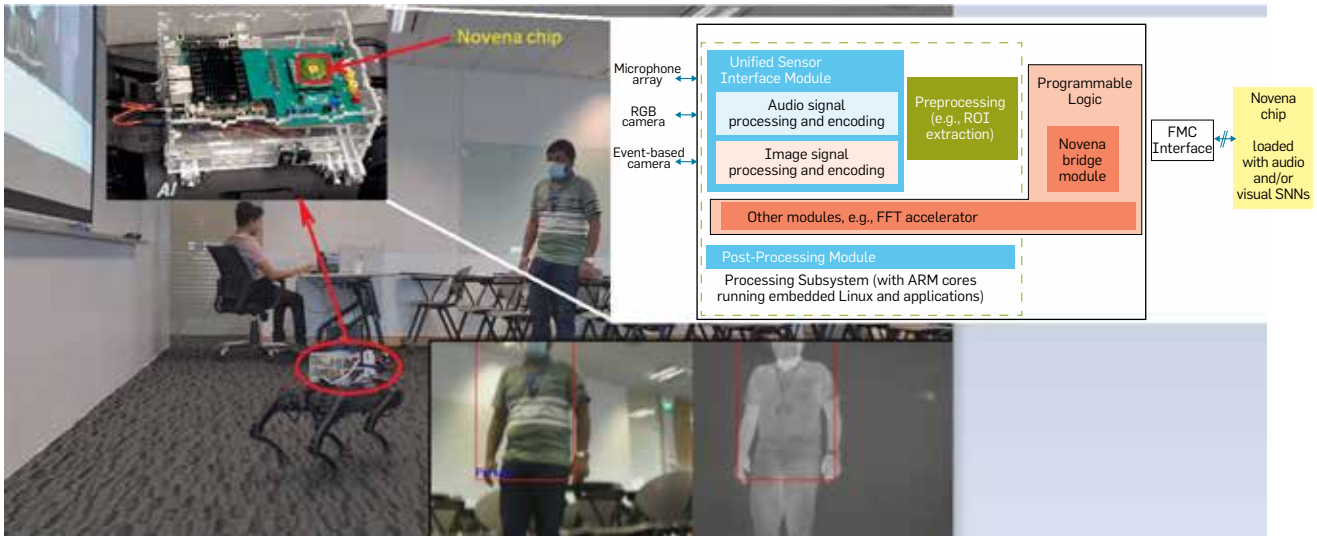
System Integration

To implement a complete neuromorphic computing system, the neuromorphic chip needs to be integrated with a host processor, along with the required sensors. We have chosen to use an FPGA board with ARM cores as the host processor.

First, the FPGA's programmable logic (PL) is programmed with a bridging module to interface with the Novena chip over an FPGA Mezzanine Card low-pin count connector. The PL may also contain other modules such as FFT library implementation to accelerate certain pre-processing steps that would otherwise take considerable cycles in the ARM cores.

Second, the ARM cores, as part of the FPGA processing subsystem, run embedded Linux, which facilitates sensor integration thanks to the availability of device drivers for a wide range of commercial sensors including some event-based sensors. The sensors are then consolidated under a unified sensor interface module, followed by further pre-processing as necessary. One example of pre-processing is to perform region-of-interest (ROI) extraction such as on event-based visual inputs,[1] so as to reduce input dimension and allow

Figure 3. Example neuromorphic computing system using a legged robot as mechanical platform and thermal + RGB camera-based human detection.



resource-efficient implementation on the neuromorphic chip.

Furthermore, a system software programming framework based on a Xilinx SDx environment has been developed to streamline communications between embedded Linux and FPGA PL and to facilitate interfacing with Novena from embedded Linux applications, so that the developer need not be overwhelmed by the low-level details of the Novena bridge interface module in FPGA PL.

Once the Novena chip returns computed outputs, a post-processing module may be used, for example, to further filter outputs over time and improve final accuracy. An illustration of the overall neuromorphic system diagram is shown in the white inset of Figure 3.

Finally, to showcase the capabilities of neuromorphic computing, several demonstrator applications have been built, including live keyword spotting (KWS), live gesture detection at variable distances,[1] and human plus body-part detection for search and rescue applications. Among these, KWS and human detection demos were integrated onto a legged robot, where an operator can use keywords to guide the robot, with the legged robot using a combination of thermal camera and RGB camera where thermal images were used for ROI extraction based on hotspots, followed by human and body-part classification on associated RGB image ROI patch. A snapshot of the human detection demo is shown in Figure 3, along with a zoomed-in view of the FPGA board and Novena chip.

Conclusion

This article gives a description of the six-year effort to develop an end-to-end neuromorphic computing solution in Singapore. From next-generation memory devices, chips, algorithms, middleware, to applications, this ambitious program covered all aspects of neuromorphic computing. Currently, the program is looking into the commercialization of its innovations, especially for energy-efficient edge AI.

Acknowledgment. This work was supported by the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain) under Grant A1687b0033. C

References

1. Fabien, C. et al. Event-based visual sensing for human motion detection and classification at various distances. In *Proceedings of Pacific-Rim Symp. Image and Video Technology*, 2022.
2. Lee, M.K. F. et al. A system-level simulator for RRAM-based neuromorphic computing chips. *ACM Trans. Architecture and Code Optimization 15*, 4 (2019), 1–24.
3. Luo, T. et al. An FPGA-based hardware emulator for neuromorphic chip with RRAM. *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems 39*, 22 (2018), 438–450.
4. Nambiar, V.P. et al. Energy efficient 0.5 V 4.8 pJ/SOP 0.93 W leakage/core neuromorphic processor design. *IEEE Trans. Circuits and Systems II: Express Briefs 68*, 9 (2021), 3148–3152.
5. Wong, M.M. et al. A 2.1 pJ/SOP 40nm SNN accelerator featuring on-chip transfer learning using Delta STDP. In *Proceedings of the IEEE 51st European Solid-State Device Research Conf.* 2021.
6. Wu, J. et al. Progressive tandem learning for pattern recognition with deep spiking neural networks. *IEEE Trans. Pattern Analysis and Machine Intelligence, 44*, 11 (2021), 7824–7840.
7. Yang, L. et al. Coreset: Hierarchical neuromorphic computing supporting large-scale neural networks with improved resource efficiency. *Neurocomputing 474* (2022), 128–140.
8. Zhang, M. et al. Rectified linear postsynaptic potential function for backpropagation in deep spiking neural networks. *IEEE Trans. Neural Networks and Learning Systems 33*, 5 (2021), 1947–1958.

Tao Luo is a research scientist at the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), in Singapore.

Weng-Fai Wong is an associate professor in the School of Computing (SoC), at National University of Singapore (NUS), Singapore.

Rick Siow Mong Goh is a research scientist at the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), in Singapore.

Anh Tuan Do is a research scientist at the Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), in Singapore.

Zhixian Chen is a research scientist at the Institute of Microelectronics (IME), Agency for Science, Technology and Research (A*STAR), in Singapore.

Haizhou Li is a professor in the School of Computing (SoC), at National University of Singapore (NUS), Singapore.

Wenyu Jiang is a research scientist at the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), in Singapore.

Weiyun Yau is a research scientist at the Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), in Singapore.

DOI:10.1145/3589736

BY MASARU KITSUREGAWA, SHIGEO URUSHIDANI, KAZUTSUNA YAMAJI, HIROKI TAKAKURA, ICHIRO HASUO, IMARI SATO, FUYUKI ISHIKAWA, ISAO ECHIZEN, AND KENSAKU MORI

Activities of National Institute of Informatics in Japan

THE NATIONAL INSTITUTE of Informatics (NII) is a national-level research and services institute focusing on informatics in Japan. NII was formally established in 2000 by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT). Located in downtown Tokyo, approximately 80 permanent researchers conduct computer science research in various fields. In addition, NII provides networking and information services for more than 1,000 research and educational institutions in Japan. This article describes NII's activities in these two complementary missions: research and services that support research (as illustrated in Figure 1).

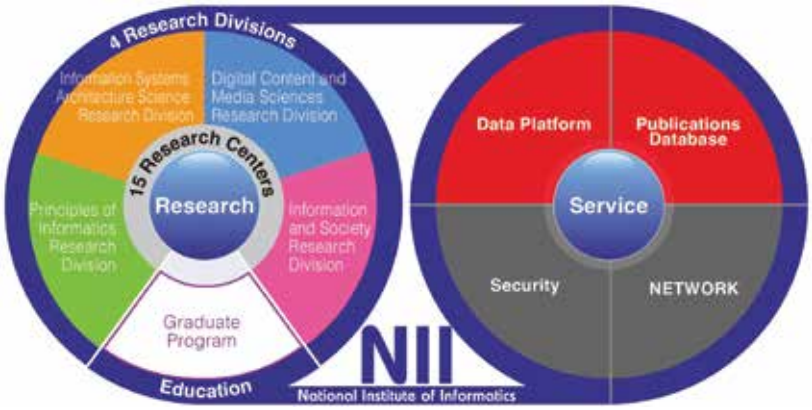
The service mission of NII has been fulfilled through three widely deployed academic services: the SINET family of high-performance and availability networks; the Gakunin Research Data Management (RDM) Platform, which supports open access to, and secure sharing of, data-driven research results; and cybersecurity services to protect SINET, RDM, and other services such as secure computation. The research mission of NII can be represented by four fundamental computer science projects and two applied research projects. The fundamental research projects are: graph algorithms, formal verification and their applications, machine vision using fluorescence, and engineerable AI. The applied research projects are: detection of fake videos, audio clips, and documents; and CT image AI-based analytics for the COVID-19 pandemic. The three groups of projects, services in support of research, fundamental research, and applied research, will be briefly described in this article.

Academic Services in Support of Research

SINET backbone network service. One of the most impactful IT services provided by NII is SINET6,[6] which has a fully redundant 400Gbps backbone network from Hokkaido to Kyushu, as well as a 200Gbps connection to Okinawa. SINET6 provides high-reliability and high-performance backbone network service to more than 1,000 research and educational institutions in Japan. SINET5 was the previous-generation network, which had a 100Gbps backbone, with an automated, very fast (less than 50msec) switch over when needed. Since SINET5 became operational, NII has provided reliable network services through many natural disasters that have hit Japan over the years.

The typical unique features of SINET6 are its high speed, high density, and multiplicity of network services. SINET6 has 70 SINET nodes nationwide and connects them by the

Figure 1. NII organizational structure.

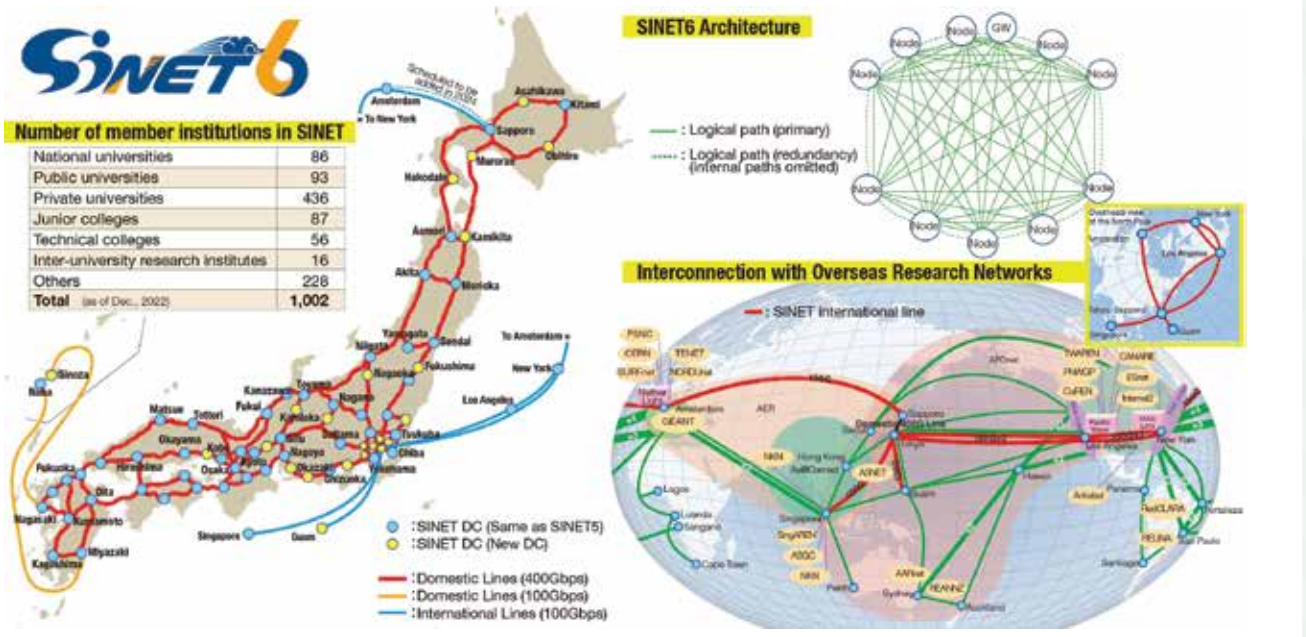


world's fastest 400 gigabit Ethernet (400GE) lines. This makes it easy for the user institutions to introduce high-speed access lines to SINET nodes at low cost. It is safe to say that SINET6 provides them with the world's fastest end-to-end network as compared to commercial networks and other research networks. SINET6 also provides a variety of network services compared with other networks. It provides both IPv4/IPv6 dual-stack services for open network access and layer-2/3 virtual private network (VPN) services for secure collaborations. SINET users can set up L2VPNs by specifying the destinations and routes in an on-demand manner through a SINET on-demand controller. This on-demand capability is useful for various research fields such as telesurgery experiments that need assured bandwidth, non-compressed 8K video transmissions that need to select routes to obtain enough bandwidth, and so on. NII has stably provided this variety of services over the cutting-edge high-speed devices by effectively developing and debugging new functions.

The development of the SINET family of networks has been led by Shigeo Urushidani, deputy director of NII. In addition to the backbone network, NII also provides Mobile SINET service for research and innovative IoT applications, such as precision agriculture (for example, cattle monitoring) and atmospheric phenomena (for example, polar mirage monitoring). In parallel with the technological evolution of the SINET family of backbone networks, Mobile SINET has evolved from LTE to 5G, with support for private 5G service (see Figure 2).

Data platform service. NII RCOS (Research Center for Open Science and Data Platform),[5] under the direction of Kazutsuna Yamaji, is responsible for the development of the NII Research Data Cloud. The initial version of the NII RDC consists of three primary platforms: a research data management platform (GakuNin RDM), a publication platform (JAIRO Cloud), and a discovery platform (CiNii Research). These platforms aim to be utilized along with the research data life cycle. GakuNin RDM is currently used by more than 60 research institutions. It aims to be a vital service that enables researchers to share data between collaborators in their daily activities. In addition to data sharing, GakuNin RDM also supports the Data Management Plan enforced by the funding agency, metadata management, and a data analysis environment with appropriate secure storage capacity. Reliable research output managed by GakuNin RDM is publicized from the cloud service of institutional repositories

Figure 2. SINET 6: Academic backbone network in Japan.



named JAIRO Cloud, which is employed by more than 750 universities in Japan. The whole contents stored in each repository are aggregated by CiNii Research which has an enormous academic knowledge graph in Japan.

In addition to the symbolic European Open Science Cloud, similar activity can be seen globally, such as Australian Research Data Commons, Korean Research Data Platform, Malaysia Open Science Platform,[7] and African Open Science Platform. NII RDC continues its effort to build interoperability with these platforms. Compared with them, the advantage of NII RDC is the national-scale deployment and comprehensive service through the research data life cycle. The National Research and Education Network (NREN) in each country provides a high-speed network and several middleware services, including identity and access management federation. The uniqueness of NII is to add value through the research data infrastructure on top of these existing services.

In 2020, the Japanese Cabinet Office recommended researchers to use NII RDC in large-scale publicly funded research projects, which is developing good use cases in practice. For instance, in the Moonshot Goal 2 project in Japan realization of ultra-early disease prediction and intervention by 2050, life scientists and theoretical scientists share and manage research data using GakuNin RDM. They also utilize data analysis tools belonging to GakuNin RDM and share analysis programs and mathematical models. The program aims to publicize their research data using JAIRO Cloud. NII RDC has become a vital research infrastructure in their research program. In 2022, Japan's MEXT started a new program for nationwide and institutional wide deployment of GakuNin RDM. Through this program, the laboratory in each university will get an opportunity to be in touch with GakuNin RDM and gain its benefits.

NII RDC is not only to provide a feasible research environment but also to change the conservative research culture to facilitate the data-centric science. As illustrative example, visualizing the adoption of a standard dataset by many research papers will increase the recognition of its impact. By adding a new dimension to the current exclusive emphasis on paper citation counts, scientists who collect and publish high-quality datasets will also become highly regarded (see Figure 3).

Cybersecurity service. Under the direction of Hiroki Takakura, the Center for Strategic Cyber Resilience R&D coordinates and supports NII Security Operation Collaboration Services (NII-SOCS), which operates security monitoring and provides technical advice for national universities in Japan. In addition to countering and mitigating overt cyberattacks, the center implements technologies for detecting signs of advanced persistent threats (APT) and deploying countermeasures to block the APT and their follow-up attacks such as ransomware.

With the advent of COVID-19, our lifestyles have been drastically transformed. This transformation also affected communications between SINET and the Internet, which are monitored by NII-SOCS. It is natural for everyone to take full advantage of the convenience of cloud computing, and in addition to legitimate users, attackers also use major cloud service providers. More than 90% of communications are now encrypted. As the results, traditional security measures, including pattern-matching, indicators of compromise, are rapidly losing their effectiveness.

Under such circumstances, we developed various analysis methods, for example, the combination of threat intelligence analysis and parallel processing by GPUs to find malicious activities in communications at several hundred Gbps.[2] As a result, we have succeeded in narrowing down the billions of suspicious commu-

Figure 3. Architecture of NII Research Data Cloud.

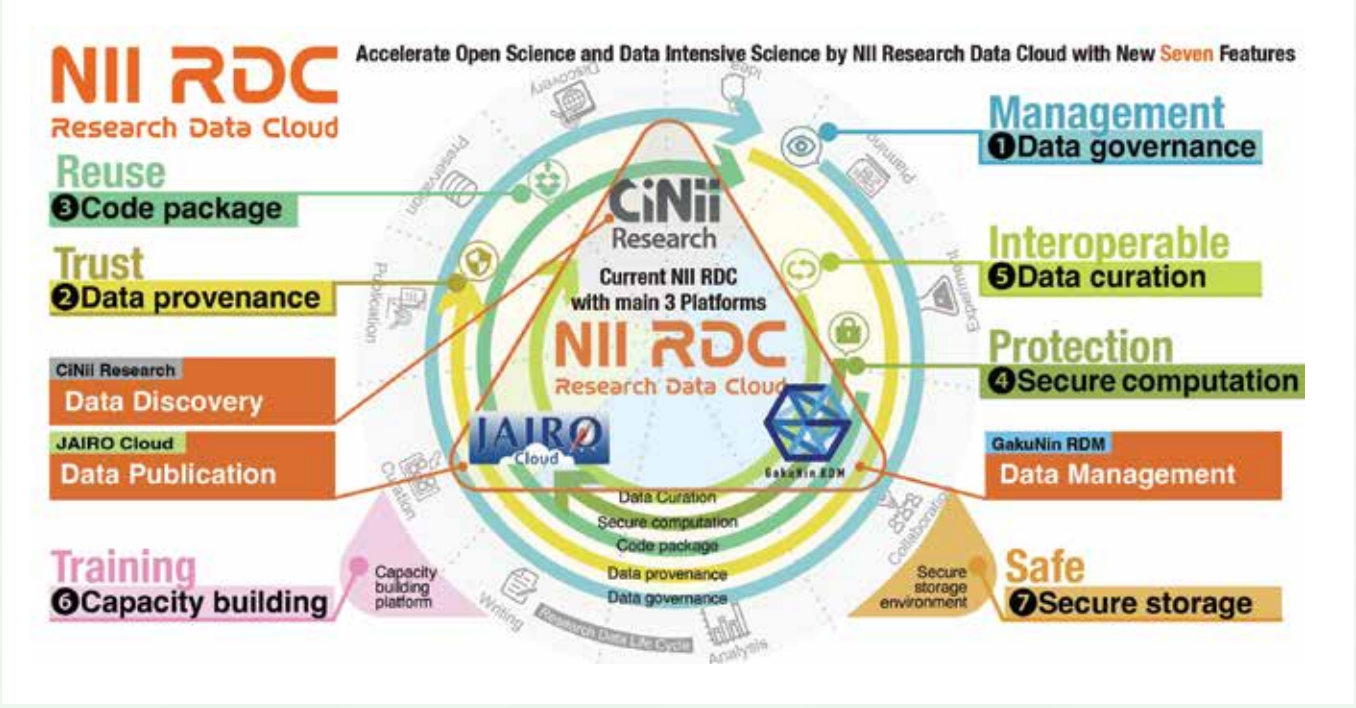


Figure 4. The pull-over scenario for autonomous driving.

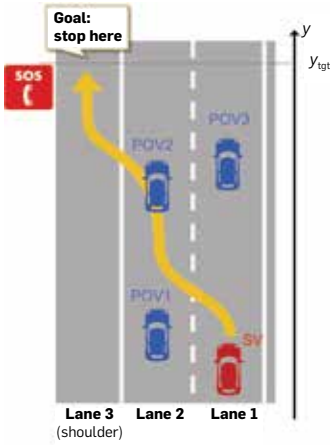
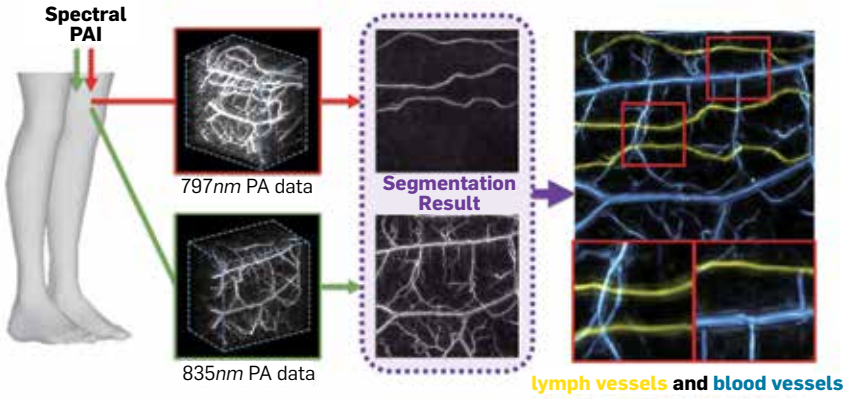


Figure 5. Photoacoustic imaging for vessels.



nications generated daily to a few dozen high-risk ones. NII-SOCS also analyzes a variety of threat information and requests enhanced defenses when vulnerabilities are discovered in the networks of participating universities. This activity contributes to the prevention of damage such as unauthorized intrusion into VPN routers, which has been recently occurring around the world.

NII-SOCS fully utilizes the project of SINET 6, traffic mirroring functions. In this service, the traffic of their access lines is mirrored and forwarded to specified hub locations with security analysis devices. Since the amount of mirrored traffic is enormous, the target for mirroring is dynamically selected and turned on or off.

To stimulate research activities on cybersecurity, the center provides research data, including sanitized traffic benchmark and malware data with analysis reports based on observed attacks.

The center also conducts table-top exercises, role-based discussion groups with scenarios modeled on actual cyberattacks that have occurred recently in universities and other organizations. These discussion-based simulated exercises modeled on actual cyberattack scenarios are designed for all enterprise personnel, including technical and administrative staff, as well as executives. These simulated exercises help universities build resilience, to continue their operations even in the face of cyberattacks. Through these activities, the center contributes to the broadening of cybersecurity expertise through the entire enterprise, providing a potential remedy for the global shortage of dedicated cybersecurity professionals.

Fundamental Research in Computer Science

Graph algorithms and challenge for minimum cut(mincut). Ken-ichi Kawarabayashi shared Fulkerson Prize in 2021 with Mikkel Thorup of the University of Copenhagen, the highest honor in discrete mathematics and algorithms, for their deterministic mincut algorithm for a simple graph. Mincut is one of the central problems in graph algorithms, starting from the seminal algorithm of Ford and Fulkerson (1956). It has been still a challenge though, to extend to *multiple* graphs.

Formal verification and applications. Ichiro Hasuo's group made significant contributions to software science, ranging from mathematical foundations to real-world applications. In their recent work[3] on safety of automated driving vehicles (ADVs), they introduced the first formal logic that proves safety in the pull-over scenario (Figure 4). The scenario, while essential for level 4–5 ADVs, poses a unique challenge due to its complexity. Their logic overcame the complexity by *compositional reasoning*, a central paradigm of deductive verification à la Antony Hoare.

Innovative computer vision based on fluorescence and photoacoustic

Figure 6. Engineerable AI for autonomous driving.

Risk analysis-driven, fine-grained safety requirements for prediction performance in safety-critical systems

ID	AI Error Modes	Scenes	Hazards	Risk Level	AI Eval.	Accepted?
001	Misclassify: pedestrian to rider	Pedestrian in front of ego-car	No braking, hit the pedestrian	5	Error 4%	Y
002	Misclassify: rider to pedestrian	Closely following car	Unnecessary braking, hit at the rear side	3	Error 35%	N
...	...	...	...	...	...	...

Targeted and regression-controlled update of deep neural networks via fault localization to identify responsible/suspicious parameters

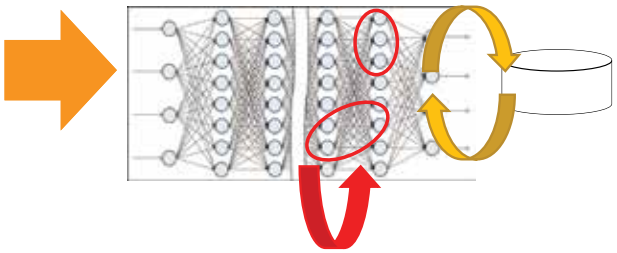
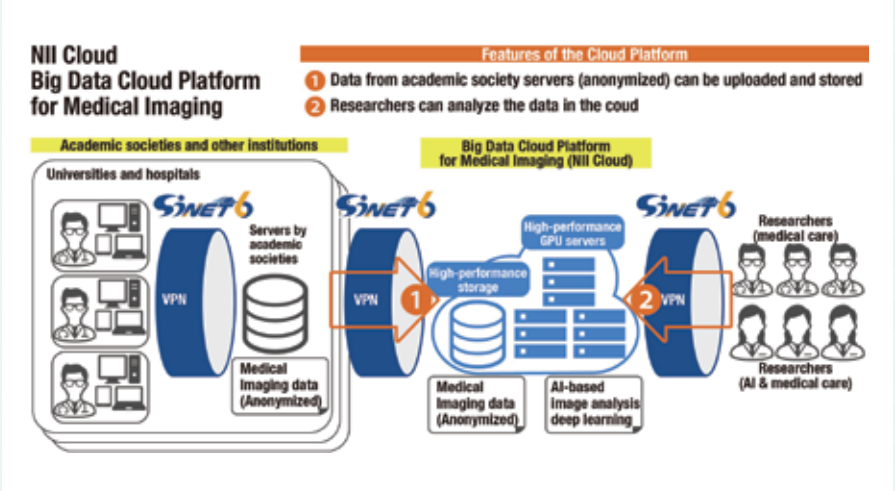


Figure 7. Medical imaging big data platform built by NII.



signals. Imari Sato's pioneering work analyzing reflective-fluorescence images has opened a new research field in computer vision.[7] Based on the unique properties of fluorescence, she presented a series of solutions to challenging computer vision problems, including light color estimation, robust shape estimation of concave objects, and liquid classification. She has also made important achievements in photoacoustic imaging technology, a state-of-the-art noninvasive measurement technique, and realized a fully automated system for 3D visualization of blood and lymphatic vessels (see Figure 5).

Engineerable AI. Fuyuki Ishikawa of NII is leading the "Engineerable AI (eAI)"[a] project for trustworthy AI engineering. One of the notable outcomes is techniques for targeted and regression-controlled update of neural networks by adapting fault localization techniques, originally used for estimating the bug locations. These techniques were tested by benchmarks defined with the automotive industry and demonstrated the capability to adjust and improve prediction performance of state-of-the-art prediction AI to meet fine-grained safety requirements for automated driving (see Figure 6).

Applied Research for Societal Benefits

Detection of fake images, audio clips, and documents. During the COVID-19 pandemic, fake news items have caused significant harm, a problem recognized as an "infodemic" by the World Health Organization (WHO). Attackers can generate and distribute fake videos, fake audio clips, and fake documents for purposes of fraud, propaganda, and public opinion manipulation, using widely available tools such as the DeepFake[1] image generator. To detect fake media, Isao Echizen and his colleagues founded the Global Research Center for Synthetic Media (SynMedia Center). The SynMedia Center conducts research and development to generate and detect synthetic and fake media, ensure media reliability, and support decision making. A concrete example is the AI-based SYNTHETIQ VISION tool (September 2021) that automatically detects fake facial videos generated by tools such as DeepFake. This fake detector program is available as a Web API and illustrates "AI as a service (AIaaS)."

AI-based COVID-19 pneumonia diagnosis tool using CT image analytics.[b] When the COVID-19 pandemic began in early 2020, PCR (polymerase chain reaction) testing stations were relatively rare in Japan. For more serious cases that required hospitalization, NII and the Japan Radiological Society developed an AI-based COVID-19 pneumonia diagnosis tool, using computerized tomography (CT) images. Although the data volume of each CT image set is very large, we were able to leverage SINET6 to collect more than 700 sets of CT images of COVID-19 pneumonia cases from around Japan. The AI-based diagnostic tool was developed quickly and achieved a detection accuracy of 83%. This achievement was reported widely, including in an article in *MIT Technology Review* (July 2021). This system was one of the earliest AI-based COVID-19 pneumonia diagnosis detectors in the world. Kensaku Mori of Nagoya University played a leading role in the NII development team, in collaboration with medical imaging researchers from the University of Tokyo, Kyushu University and others, achieving success through Team Science (see Figure 7).

Conclusion

A growing number of successful NII projects have built on vigorous international collaborations. For example, the 2021 Fulkerson Prize was given in recognition for a paper by Kawarabayashi and Thorup of the University of Copenhagen. Beyond the illustrative projects described in this article, the NII team sincerely wishes to expand global collaborations in many areas of computer science research. C

a See eAI Project; https://engineerable.ai/en/

b See http://research.nii.ac.jp/rc4mb/

References

1. Afchar, V. et al. MesoNet: A compact facial video forgery detection network. In *Proceedings of the 2018 IEEE Intern. Workshop on Info. Forensics and Security*, 1–7; doi:10.1109/WIFS.2018.8630761.
2. Ando, R., Kadobayashi, Y. and Takakura, H. Choice of parallelism: Multi- GPU driven pipeline for huge academic backbone network. *Intern. J. Parallel, Emergent and Distributed Systems 36*, 4 (2021), 6.
3. Hasuo, I. et al. Goal-aware RSS for complex scenarios via program logic. *IEEE Trans. Intelligent Vehicles*, 2022; doi:10.1109/TIV.2022.3169762.
4. Kawarabayashi, K. and Thorup, M. Deterministic edge connectivity in near-linear time. *JACM 66*, 1 (Dec. 2018). Article 4, 1–50, https://doi.org/10.1145/3274663
5. Komiyama, Y. and Yamaji, K. Nationwide research data management service of Japan in the open science era. In *Proceedings of the 6th IIAI Intern. Congress on Advanced Applied Informatics*, 2017,129–133; https://doi.org/10.1109/IIAI-AAI.2017.144
6. Kurimoto, T. et al. SINET6: Nationwide 400GE-based academic backbone network in Japan. To appear in OFC2023.
7. Zhang, C. and Sato, I. Image-based separation of reflective and fluorescent components using illumination variant and invariant color. *IEEE Trans. Pattern Analysis and Machine Intelligence 35*, 12 (2013), 2866–2877.

Masaru Kitsuregawa is Director-General of National Institute of Informatics and University Professor at the University of Tokyo, Japan.

Shigeo Urushidani, Kazutsuna Yamaji, Hiroki Takakura, Ichiro Hasuo, Imari Sato, Fuyuki Ishikawa, and **Isao Echizen** are professors at the National Institute of Informatics, Tokyo, Japan.

Kensaku Mori is Director of the Information Technology Center, Nagoya University, and Director of NII Research Center for Medical Bigdata, Tokyo, Japan.

Responsible AI | DOI:10.1145/3589946

Operationalizing Responsible AI at Scale:

CSIRO Data61's Pattern-Oriented Responsible AI Engineering Approach

BY QINGHUA LU, LIMING ZHU, XIWEI XU, JON WHITTLE, DIDAR ZOWGHI, AND AURELIE JACQUET

FOR THE WORLD to realize the benefits brought by AI, it is important to ensure artificial intelligent (AI) systems are responsibly developed, used throughout their entire life cycle, and trusted by the humans expected to rely on them.[1] The goal for AI adoption has triggered a significant national effort to realize responsible AI (RAI) in Australia. CSIRO Data61 is the data and digital specialist arm of Australia's national science agency. In 2019, CSIRO Data61's worked with the Australian government to conduct the AI Ethics Framework research. This work led to the release of eight AI ethics principles to ensure Australia's adoption of AI is safe, secure, and reliable.[a]

This effort is influenced by Australia and New Zealand's unique culture values, such as "fair go," indigenous community, and the wider Indo-Pacific regional thinking. To test out the impact of the framework, CSIRO Data61's interviewed CSIRO scientists in 2020 to gain insight into their implementation of ethics principles in scientific projects. In 2021, the adoption of the ethics principles was examined via case studies from some of Australia's biggest businesses.[b]

It is challenging to turn high-level AI ethics principles into real-life practices.

Through our engagement with the Australian industry and the wider Oceania region, we have identified five major challenges in operationalizing RAI at scale:

- **Challenge 1.** *A lack of connected reusable solutions for operationalizing these principles.* It is challenging to turn those high-level AI ethics principles into real-life practices. Also, significant efforts have been made on algorithm-level solutions focusing on a subset of principles (such as fairness and privacy). To try to fill the gap, further guidance (such as guidebooks and checklists) has started to appear, but those efforts tend to be ad hoc sets of prompts.[2]
- **Challenge 2.** *Stakeholder interests in RAI risk management vary, demanding different but connected solutions.* Organizations usually take a risk-based approach, but different

a http://bit.ly/3FbunOj

b http://bit.ly/3kTXoaF

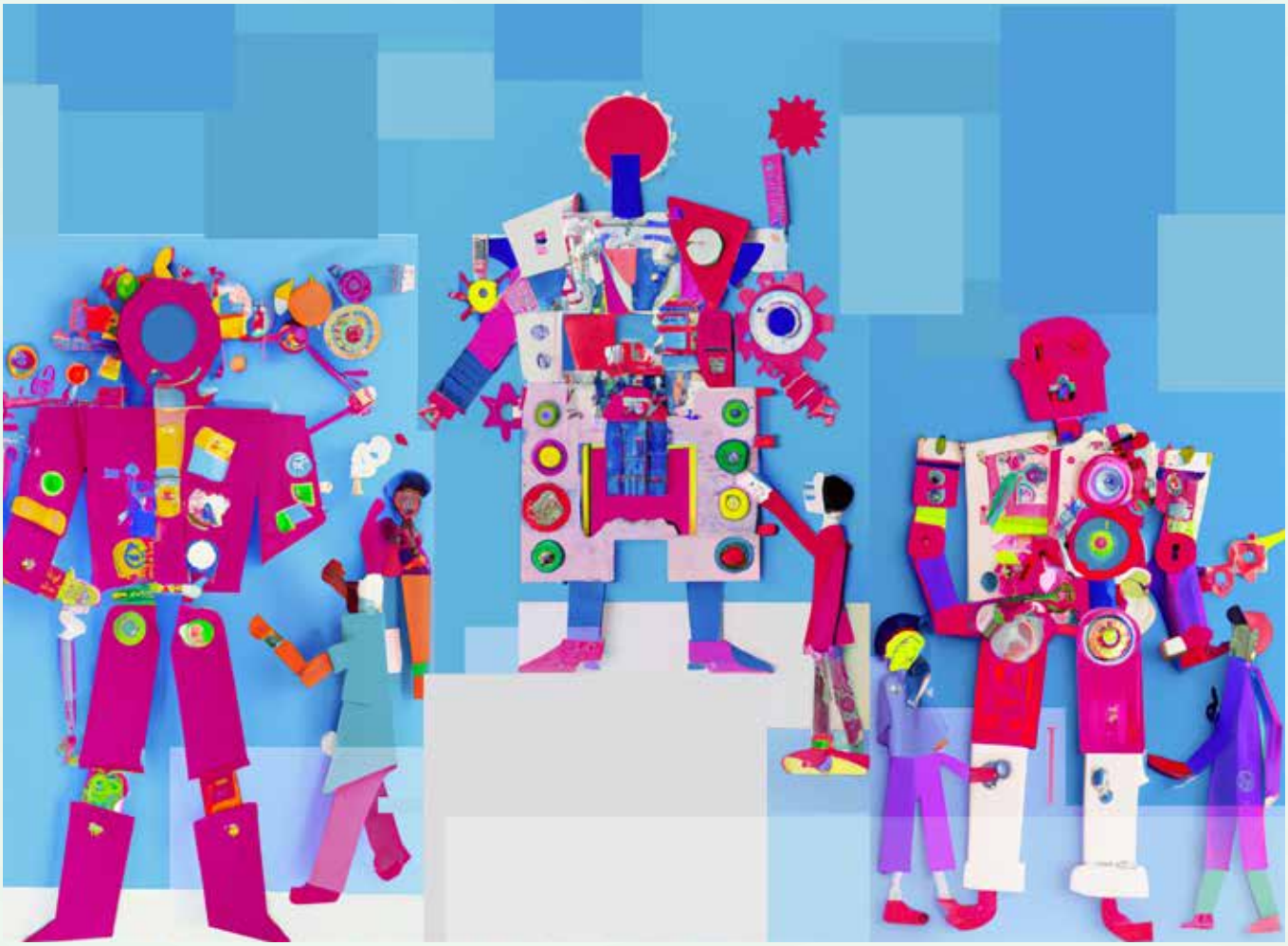


IMAGE CREATED BY ANDRIJ BORYS ASSOCIATES WITH THE ASSISTANCE OF SHUTTERSTOCK.AI

stakeholders have different interests. For example, regulators or board/executives may be more interested in harms and societal impacts, associated preventive/corrective cost, and governance mechanisms, while developers care more about algorithmic risks and reliability/security techniques.

► **Challenge 3.** *Integrating RAI risk management into organizations' existing governance frameworks.* Organizations usually have centralized risk committees to manage traditional risks, such as financial, reputation, and legal risks, with some having dedicated cybersecurity and privacy risk management. These committees do not have the necessary scope to handle RAI risks. But creating dedicated AI risk committees for human values or fairness gives rise to risk silos. Some risks, such as reliability, might be best handled at the operational level.

► **Challenge 4.** *Communication barrier.* It is often difficult for software engineers to explain the RAI risks of or solutions for their AI systems to management teams and vice versa.

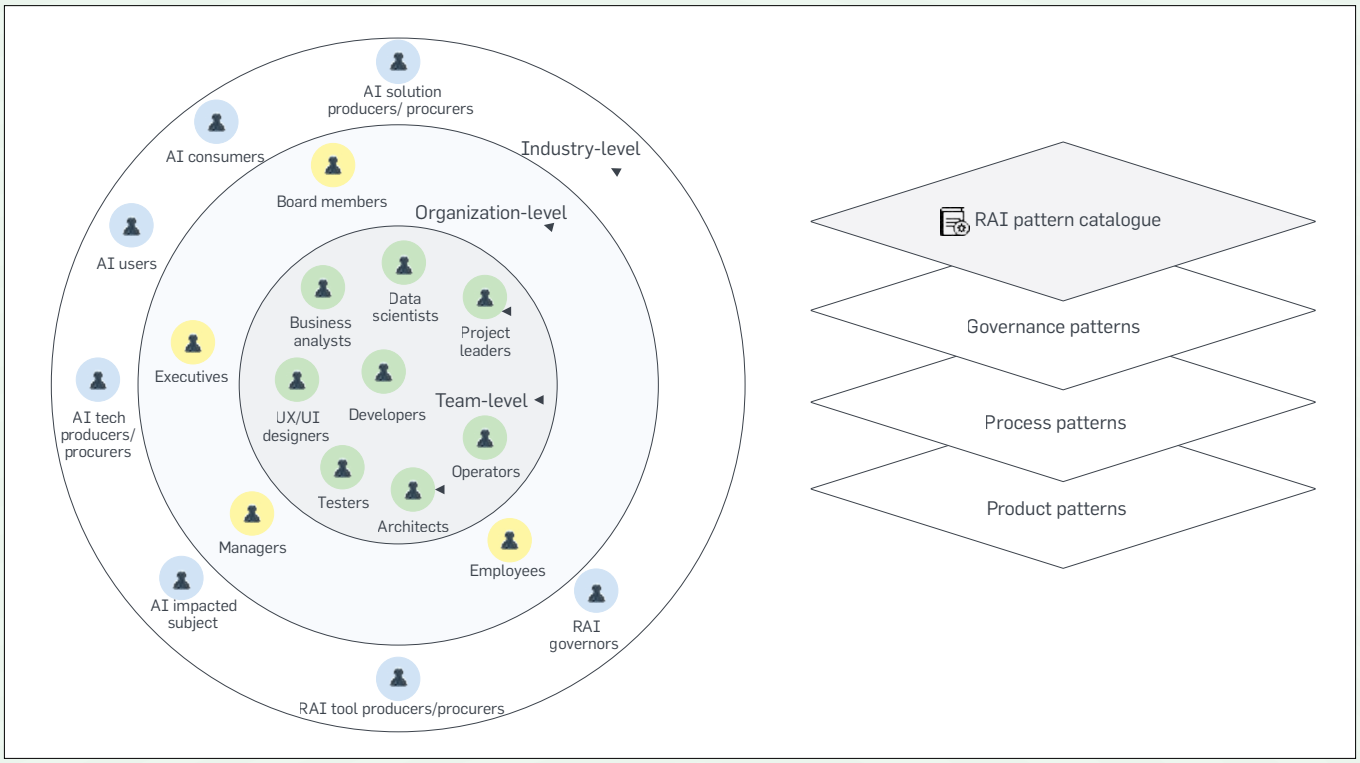
► **Challenge 5.** *Lack of talents and experts.* Organizations often do not have RAI expertise; thus their traditional risk experts manage RAI risks. Also, the organizations often cannot afford to examine each AI project deeply, thus their risk committees only assess high-risk projects and rely on the project teams to do self-assessment.

RAI Pattern Catalogue and its Research Impact

CSIRO Data61 focuses on RAI engineering, which addresses end-to-end system-level challenges. We built an RAI Pattern Catalogue[c] for different stakeholders in the AI industry.[2,3] We analyzed and generalized successful case studies and best practices into reusable patterns (Challenge 1) and organized them into three interconnected categories (Challenge 2) for easier adoption for impact—*governance patterns* for establishing multilevel RAI governance, *process patterns* for setting up responsible development processes, and *product patterns* for building RAI-

c http://bit.ly/3T0663F

CSIRO Data61's worked with the Australian government and industry to conduct research in Responsible AI Engineering.



Responsible AI Pattern Catalogue.

by-design (as illustrated in the accompany figure).

To describe the pattern, we created a template, including summary, type, objective, target users, impacted stakeholders, relevant principles, context, problem, solution, consequences, related patterns, and known uses. Patterns are selected and tailored for different contexts; for example, adding more quantitative assessment of risk residues and risk/cost increase in other parts. To maintain the risk profile for consequences, we have added pointers to measures, measuring methods, new risks, and risk residues (Challenge 3).

Patterns are connected at multilevels/stages through the related patterns (Challenge 4). For example, the regulatory sandbox is an industry-level governance pattern that may specify high ethical-quality requirements and use restrictions for facial recognition. To support the regulatory sandbox pattern, the software bill of materials pattern at the organization-level and verifiable claim for AI system artifacts pattern at the team level can be implemented to manage the ethical qualities of the procured third-party facial-recognition technology. RAI governance through the API's process pattern can be adopted to restrict the use of facial recognition APIs to approved users only. The bill of materials registry product pattern can be embedded as a system component to record the supply chain information.

RAI governance through the API's process pattern can be adopted to restrict the use of facial recognition APIs to approved users only.

In one task, we are assessing risks for CSIRO AI projects using our top-to-bottom/end-to-end RAI risk framework and recommending pattern-oriented mitigation strategies. The question bank for risk assessment is built based on four dimensions: ethics principles, life-cycle stages, stakeholders who will ask the questions, and stakeholders who will answer the questions. To support automated self-assessment, we are building a RAI knowledge graph based on incidents, questions, and patterns, among others (Challenge 5).

We offer a course on the RAI Pattern Catalogue in Australia's nationwide AI graduate schools. Our industry partners have adopted some patterns in their governance policies and development guidelines—for example, for chatbot development.[4] C

References

1. Lu, Q. et. al. Towards a roadmap on software engineering for responsible AI. In *Proceedings of CAIN 2022*.
2. Lu, Q. et. al. Responsible AI pattern catalogue: A multivocal literature review. 2022; arXiv:2209.04963.
3. Lu, Q. et. al. Responsible-AI-by-design: A pattern collection for designing responsible AI systems. *IEEE Software* (2023).
4. Lu, Q. et. al. Developing responsible chatbots for financial services: A pattern-oriented responsible AI engineering approach. *IEEE Intelligent Systems* (2023).

Qinghua Lu, Liming Zhu, Xiwei Xu, Jon Whittle, Didar Zowghi, and **Aurelie Jacquet,** Data61, CSIRO, Australia.

Data Integrity | DOI:10.1145/3589154

The Veracity Grand Challenge in Computing:

A Perspective from Aotearoa New Zealand

BY MARKUS LUCZAK-ROESCH, MATTHIAS GALSTER, AND KEVIN SHEDLOCK

THE NEW ZEALAND government identified numerous challenges related to trust and truth in the context of digital technologies. These challenges result from an ever-increasing amount of online social networks, end-to-end digital supply chains, automated decision-making tools, generative artificial intelligence (AI), and cyber-physical systems. Such challenges impact people's lives across professional and private contexts and led to the *Veracity Project*[a] 2021–2024.

But what is *veracity*? Outside the field of computing, veracity is not a common term in everyday language. One dictionary definition is "conformity with truth or fact."[b] But does that definition really capture the entire meaning of veracity within computing?

Veracity was initially introduced as the fourth "V" to Big Data's original three Vs—volume, velocity and variety—often interpreted as data quality.[3] However, dictionaries also define veracity as the "power of conveying or perceiving truth," highlighting a sender-receiver perspective.

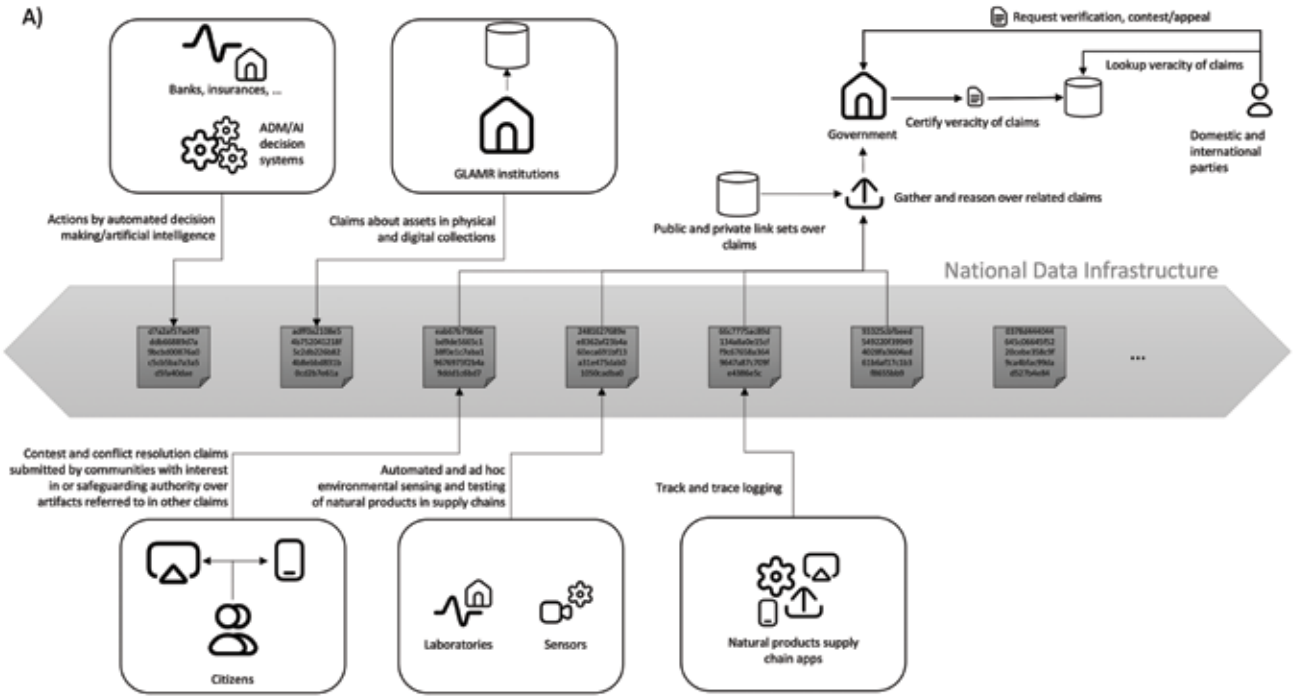
IMAGE BY MASTOCKER

Conceptualizing veracity as data quality does not capture its scope multidirectionally; and many low-trust scenarios arise from someone accessing high-quality data but doing something with that data that harms others. An example is the Facebook-Cambridge Analytica scandal, a low-veracity situation involving high-quality data.

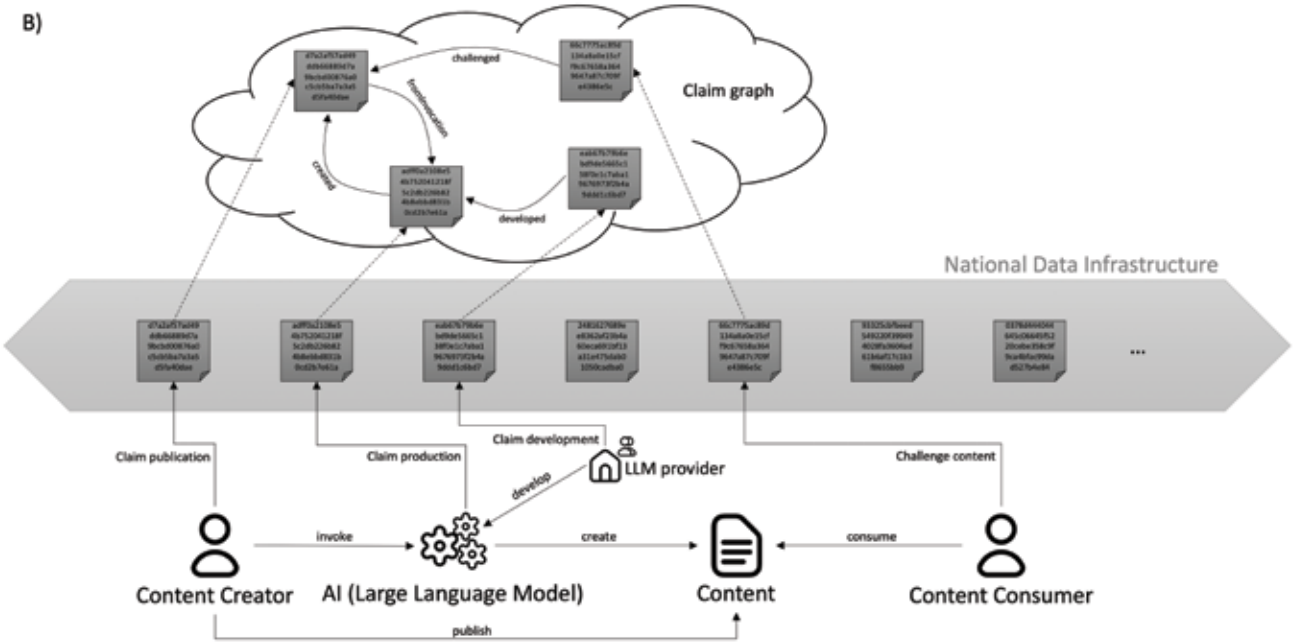
Transcending boundaries between physical and digital spheres. This conceptual ambiguity makes veracity uniquely interesting within Aotearoa New Zealand and other countries where multidirectional considerations around sovereignty of artifacts (including digital artifacts) are culturally and legally embedded in society. New Zealand is an example of a bicultural society with a commitment to recognize language and culture of the Indigenous peoples, Māori. Technology, and how technology is developed and used, must be culturally responsive and inclusive. This is specifically visible in requirements around data sovereignty. Research from Aotearoa New Zealand can move beyond common thinking about computation by fostering a co-design process that allows Indigenous and global perspectives to jointly create this new conceptualization of veracity.

Led by the affiliated Indigenous researchers, the Veracity Project research began from the notion of "Ko te taiao matihiko," emphasizing a holistic view to the world around us, and dissolving boundaries between the physical and digital worlds. This provides a novel perspective on computing in which digital representations of artifacts cannot be disentangled from their grounding in the physical world. This grounding could be either the real-world context of a digital artifact[5] (for example, the person that used a digital service is connected to the data, and the place in which the app was used) or the real-world artifact itself about which data is created (for example, an apple

a https://veracity.wgtn.ac.nz/
b https://www.merriam-webster.com/dictionary/veracity



A) To ensure and prove veracity, people, organizations (for example, from the GLAMR sector including galleries, libraries, archives, museums and record-keeping institutions), and applications (including automatic decision making/ADM systems) preregister secure identifiers (for example, embodying time and digital signatures) related to claims they want to make while maintaining full control over the actual data related to the claim. Claims can be about a certain transaction like the exchange of an artifact between parties, or about the condition or properties of an artifact at a certain time. Witness accounts about the truthfulness of claims as well as contesting of claims can also be submitted. All claims are linked by context and can be examined to assess the state of veracity.



B) Example instantiation of the architectural concept in the context of AI-supported content authoring involving a Large Language Model (LLM). Each actor in the process (including the AI system) submits claims related to their involvement in the content production, publishing and consumption process. A content consumer, for example, may submit a claim that is a challenge against the veracity of the produced content (for example, to flag misinformation or cultural appropriation). A claim graph can be constructed from all claims to infer the veracity of the content publication life cycle and each actors' role and stake.

Icons in figure from Slim Icon Set, MIT License; https://iconduck.com/sets/slim-icon-set.

Conceptual view of the envisioned data infrastructure approach to veracity.

traveling through a supply chain along with supporting metadata). The traditional approach in computing is to introduce digital twins as a "copy" of the real artifacts. But when entangling digital and physical representations, both exist and evolve as one.

This holistic concept of veracity is challenged by blind spots between the physical and digital spheres. Currently, there is no technology available that allows objects of arbitrary makeup (for example, a biological object like an apple) or scale (a car's individual screws) to hold data about itself, its life cycle, or the integrity placed on it by the means of production (organic) without attaching something to it (physical label, chemical nanodots). The disembodiment of data from its origin has become a pressing sociocultural, legal, and technical concern, and makes veracity one of the grand challenges in computing.

Trust, truth, authenticity, and demonstrability. The Veracity Project is investigating cultural and formal approaches to both clarify and explore veracity. Formal approaches help clarify because they abstract away from any example to get a clear view of problems, and then allow expression of that view in a rigorous language, which is equipped with rules that preserve properties (for example, trust, truth, authenticity, and demonstrability). For example, event calculus helps model supply chains, use an "engine" to explore the models to see how robust or complete the real supply chains might be, and develop a logic based on constructive logics[2] to pin down the essence of veracity.

Achieving high veracity through national data infrastructures. There are numerous situations where one party could disregard another party's interests to maximize their own benefit, thus lowering veracity. The Veracity Project is exploring use cases as diverse as tracking the use and integrity of Traditional Knowledge Labels attached to cultural heritage records;[c] supporting transparency and trust in the supply chain of organics products; and making AI decisions contestable. Stakeholder interviews indicate that veracity decomposes (at least) into technical veracity, process veracity, financial veracity, regulatory veracity, and cultural veracity.

Commonly recommended approaches like blockchain do not always support veracity as envisioned here. Veracity may involve real-world interactions; blockchain-based solutions may not capture all the information to establish veracity around those or the principles inherent in blockchains may exclude certain users. Furthermore, blockchain may help ensure the integrity of data records once they are in the system, but we still need mechanisms to ensure data records that go into a system are trustworthy. Finally, some types of blockchain may not be practical in all cases, for example, when enhancing existing systems with provenance mechanisms, or in domains where scalability and energy consumption are key drivers.

The Veracity Project works toward a novel kind of national data infrastructure involving decentralized data management following the principles of 'solid' data pods,[4] pre-registration of secure footprints of digital actions an entity (for example, a person, organization, or artificial agent) wants to be able to make claims or withstand scrutiny when challenged, and the ability for witnesses to vouch for or contest other entities' claims or actions. Related views have recently been articulated about privacy-preserving computation[1] and personal data management.[5]

The accompanying figure depicts our concept of such a national data infrastructure that focuses on open standards and interoperability at the technical level, and data ownership, consent, and contesting at the governance level. We investigate which open standards (for example, decentralized identifiers, verifiable credentials) enable infrastructures that scale, keep costs low while supporting legacy technologies, respect data sovereignty, and reduce technical dependencies.

Acknowledgments. The following team members of the Veracity Project also contributed substantially to the concepts and ideas conveyed in this article (listed in alphabetical order): Kelly Blincoe, Stephen Cranefield, Jens Dietrich, David Eyers, Brendan Hoare, Maui Hudson, Tim Miller, and Steve Reeves.

This work is supported under the New Zealand National Science Challenge Science for Technological Innovation (SfTI) spearhead project "Veracity Technology."

The disembodiment of data from its origin has become a pressing sociocultural, legal, and technical concern, and makes veracity one of the grand challenges in computing.

References

1. Agrawal, N., Binns, R., Van Kleek, M., Laine, K. and Shadbolt, N. Exploring design and governance challenges in the development of privacy-preserving computation. In *Proceedings of the ACM CHI Conf. Human Factors in Computing Systems*, May 2021, 1–13.
2. Bridges, D. and Reeves, S. Constructive mathematics in theory and programming practice. *Philosophia Mathematica 7*, 1 (1999), 65–104.
3. Rubin, V. and Lukoianova, T. Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online 24*, 1 (2013), 4.
4. Sambra, A.V. et al. Solid: A platform for decentralized social applications based on linked data. MIT CSAIL & Qatar Computing Research Institute, Tech. Rep., 2016.
5. Shedlock, K. and Vos, M. A conceptual model of Indigenous knowledge applied to the construction of the IT artefact. In *Proceedings of the 31st Annual CITRENZ Conf.* July 2018, 1–7.
6. Verbrugge, S., Vannieuwenborg, F., Van der Wee, M., Colle, D., Taelman, R., and Verborgh, R. Towards a personal data vault society: an interplay between technological and business perspectives. In *Proceedings of the 60th FITCE Communication Days Congress for ICT Professionals: Industrial Data–Cloud, Low Latency and Privacy*. IEEE, Sept. 2021, 1–6.

Markus Luczak-Roesch is an associate professor of information systems in the School of Information Management at Victoria University of Wellington, New Zealand.

Matthias Galster is an associate professor of computer science and engineering at the University of Canterbury, Christchurch, New Zealand.

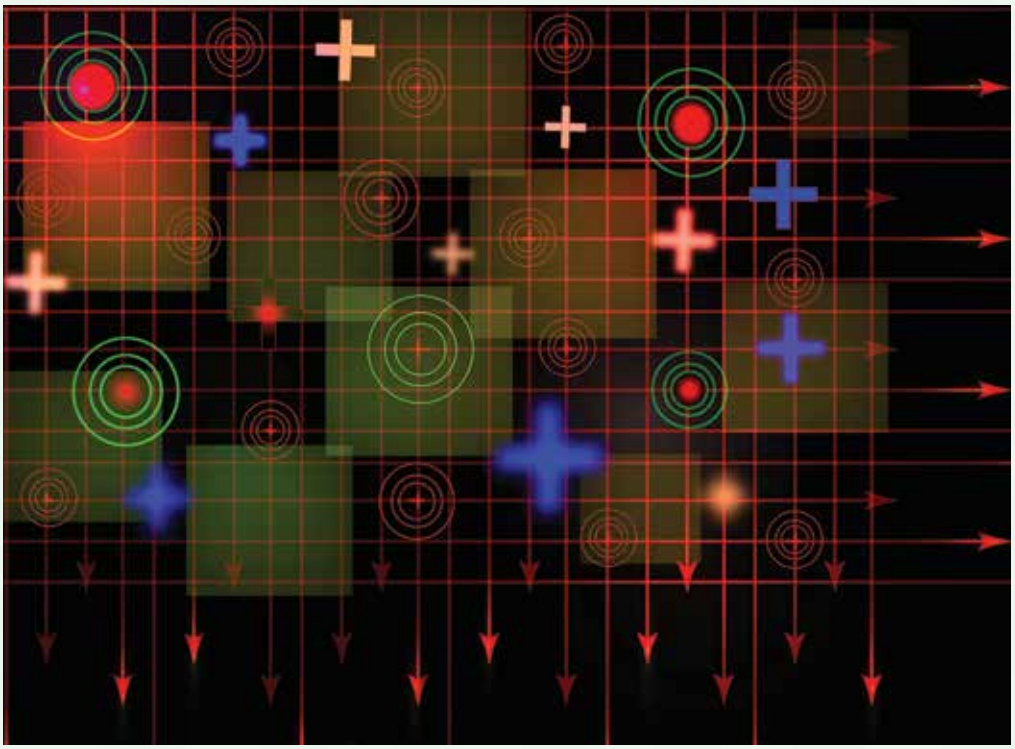
Kevin Shedlock (Ngāpuhi, Ngāti Porou, Whakatōhea) is an assistant lecturer in the School of Engineering and Computer Science at Victoria University of Wellington, New Zealand.

c https://www.localcontexts.org

COI Spotting | DOI:10.1145/3589151

CLOSET: Data-Driven COI Detection and Management in Peer-Review Venues

BY SOURAV S. BHOWMICK



OUR PEER REVIEW process aims to ensure our published results truly advance the state of the art. Our professional societies (for example, ACM, IEEE, VLDB Endowment) have worked hard toward ensuring the review and selection processes are fair and continue to fulfill this intended purpose. Our community has introduced various refinements such as single- and double-anonymous reviews, automated matching of reviewers to topics and to individual papers, multiple submission deadlines, opportunities for revision and rebuttal, "review quality week" to manually ensure quality of reviews, and a variety of guidelines for reviewer selection and manuscript handling. These refinements have greatly helped to improve the quality of refereed publications.

Yet during all these refinements to review processes, the operational definition of *conflicts of interest* (COI) and the means of declaring and handling them has changed little.[1,2] Venues typically define a COI as valid if two people wrote a paper together in the (recent) past, are at the same institution currently or in the near past, are close relatives or friends, are advisor and (former) advisee, or worked together closely on a project in the (recent) past. Conference venues often require authors to self-report COI with all potential reviewers—that is, the members of the program committee (PC). As conference PCs may easily swell beyond 200 members and with thousands of reviews to assign, PC chairs cannot realistically double-check such manual declarations of COI. Consequently, there is a great need for a framework that can automatically check whether all conflicts were reported, which is important in any case but vital if anyone should ever try to circumvent the system by deliberately under-reporting COI.[1,2]

CLOSET detects co-authorship and institution-based unreported COIs and COI violations based on a venue-specific COI policy.

CLOSET is a novel data-driven system, designed, and developed at Nanyang Technological University (NTU) for detecting and reporting unreported COIs and COI violations, if any, in a peer-reviewed venue.[a]

a CLOSET; http://bit.ly/3K1BiNa

IMAGE BY STOCKFOREST.NET

The accompanying figure depicts the framework of CLOSET. Given a set of bibliographic data sources (for example, DBLP) and a set of reviewers (initially) assigned to each paper by a review management system (RMS; for example, Microsoft's CMT, EasyChair) hosting a specific venue, it first takes a semiautomated approach to prepare venue-specific data for COI management by integrating, cleaning, and storing relevant data. Next, it deploys an efficient COI-detection technique that automatically finds various *types* of COI.

The current version of CLOSET detects co-authorship and institution-based unreported COIs and COI violations based on a venue-specific COI policy. It also detects potential cases of *submarine* COIs[1] between author-reviewer pairs for submissions, which are COIs due to potential bias that may still exist between two people even if they do not have past co-authorship/co-worker/family relationship. Finally, these results are exposed to PC chairs in a variety of user-friendly reports (for example, structured tables, charts, network views) with natural language-based explanations to aid their decision making. Note that CLOSET is orthogonal to the "secret sauce" an RMS uses for reviewer assignment to papers. The results of CLOSET can also be uploaded to an RMS to facilitate efficient COI-free reviewer *reassignment*. Currently, CLOSET has successfully interfaced with CMT to this end.

At first glance, it may seem the COI detection and management problem is straightforward as we can simply check the existence of an author's name in the co-author list of a reviewer from any bibliographic data source or check if an author-specified domain name is identical to that of the reviewer (such as a co-worker). While this may unearth some unreported COIs, our investigation with real-world data reveals that such a strategy may often miss out many valid COI cases due to challenges brought by the characteristics of RMS data (for example, unstructured, dirty, missing data, homonymous names).

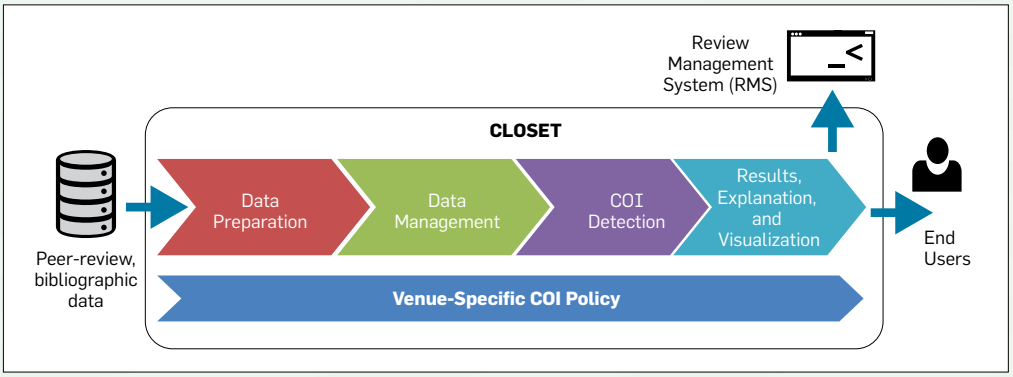
Consequently, CLOSET has an built-in data cleaning and an efficient homonymous name detection mechanism to tackle these challenges. In particular, the explanation generation component of CLOSET annotates potential homonymous name cases in results with explanations for PC chairs to investigate.

CLOSET has made a significant impact on how the review process is managed at premium data management venues.

Furthermore, automatic discovery of submarine COIs is an intriguing problem. To this end, CLOSET leverages a novel algorithm grounded on network science, data analytics, and social psychology theories (for example, social influence theory, social impact theory).

To date, CLOSET has been deployed in more than 18 venues. It revealed the COI management mechanism in existing industrial-strength RMS can often be inadequate for the intended purpose. For all deployed venues, CLOSET has detected a significant number of unreported COIs (on average, at least 25% of the submissions in a venue have unreported COIs) and COI violations that have evaded the detection mechanisms of hosting RMS. Specifically, it has made a significant impact on how the review process is managed at premium data management venues such as SIGMOD and VLDB. It has influenced changes to the specifications and implementation of long-standing COI policies and management of COI in these venues.

CLOSET is now used as the default tool in these venues. It has also inspired several community members to contribute on top of it (for example, richer visualization support by the University of Illinois Urbana-Champaign group led by Marianne Winslett, a plug-in for ingesting data from several theoretical computer science venues led by Christel Baier). In summary, CLOSET is a successful example of an end-to-end, data-driven system originating from a southeast Asian university, built on the solid foundation of network science, data analytics, and social psychology theories, and used in practice. We are not aware of any other system with similar features as CLOSET deployed in the real world. C



CLOSET framework.

References

1. Bhowmick, S.S. The end of manual checking era: It's time for automating conflicts of interest management in peer-reviewed venues. *ACM SIGMOD Blog*, May 2021.
2. Winslett, M. and Snodgrass, R. We need to automate the declaration of conflicts of interest. *Commun. ACM 63*, 10 (Oct. 2020).

Sourav S. Bhowmick is an associate professor in the School of Computer Science and Engineering at Nanyang Technological University (NTU), Singapore.

Education | DOI:10.1145/3589660

Learning and Evidence Analytics Framework Bridges Research and Practice for Educational Data Science

BY HIROAKI OGATA, RWITAJIT MAJUMDAR, AND BRENDAN FLANAGAN

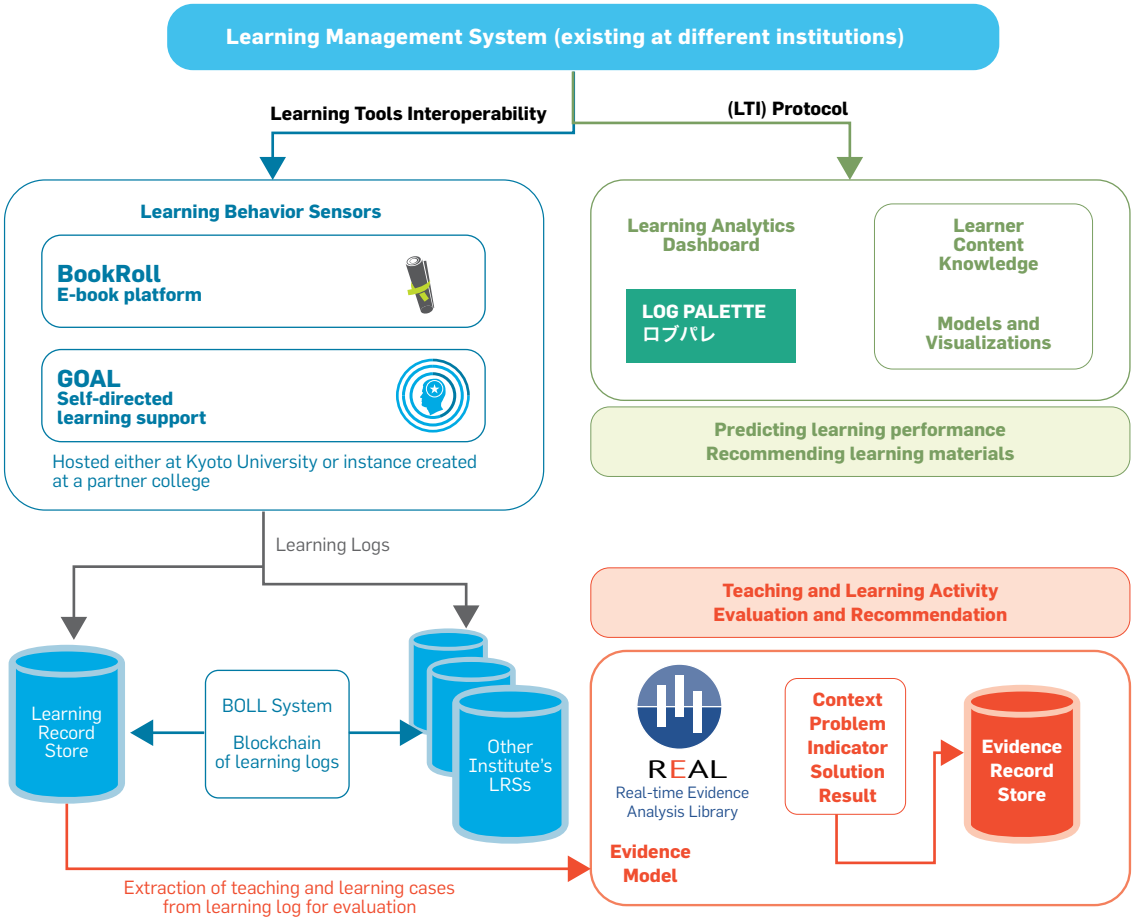
LEARNING ANALYTICS (LA) as a research discipline focuses on multiple perspectives of understanding and supporting educational activities utilizing collected log data. To do so at a national and even international level, educational technology platforms that enable gathering users' interaction traces and digitally generated artifacts must store data in a standardized format. In Japan, the government initiated the GIGA School project in 2020, which installed more than nine million tablet PCs and high-speed Internet access at compulsory education institutions (elemental and middle schools). Such infrastructure enables the collection of educational data and analysis with the aim to improve educational practices in each school. With standardized data logging, it is possible to aggregate data from all schools and to generate educational Big Data that can support evidence-based policy-making and research at a national level.

PHOTO BY KPG-PAYLESS

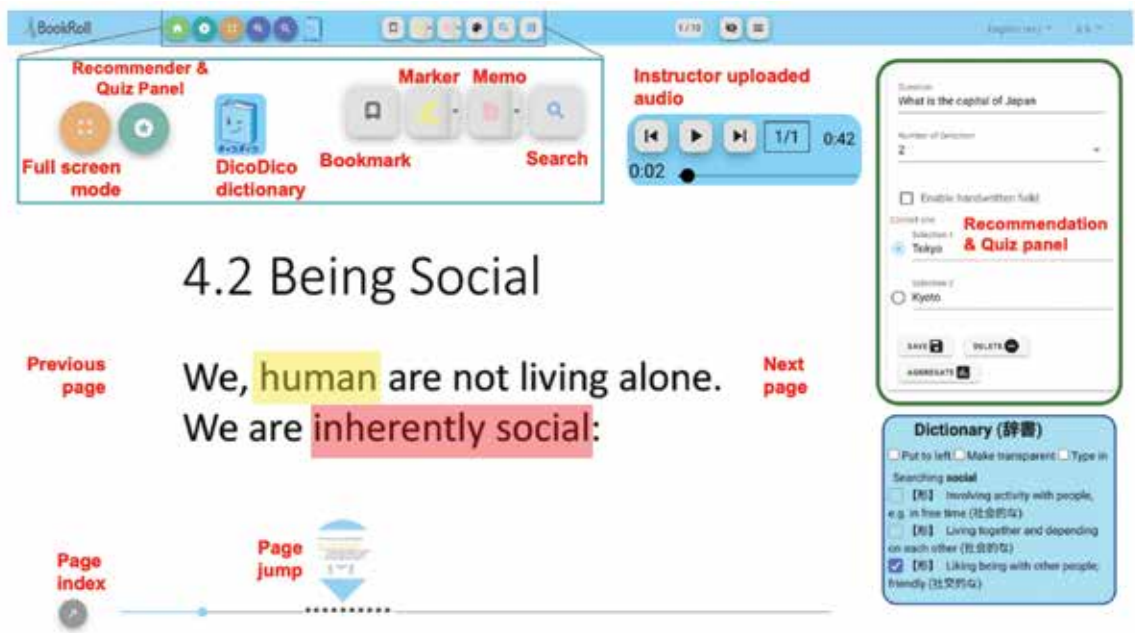
LEAF is an integrated technology framework that incorporates learning analytics methods and tools and can be linked to any existing learning management systems in different institutions.

At the Learning and Educational Technology Research Unit at Kyoto University, we developed the Learning and Evidence Analytics Framework (LEAF),[1] an integrated technology framework that incorporates LA methods and tools and can be linked to any existing learning management systems in different institutions. The accompanying figure provides an overview of the framework. As illustrated in the figure's Part A, teachers and learners can conduct daily educational activities and access various AI-driven services utilizing the LEAF system to enrich their learning experiences and outcomes. Teachers can upload any learning materials in PDF format on BookRoll, the e-book platform, and learners can access them through standard Web browsers on their PC or mobile devices (as shown in Part B). The learning interactions are logged using standard xAPI statements and then processed to create various data-driven services in LEAF. It is then presented to the users within BookRoll or in a learning dashboard called Log Palette.

Learning and Evidence Analytics Framework (LEAF).



A. Enabling educational data science practice and research.



B. BookRoll interface and the different affordances for learning and teaching.

The five tenets supporting data-driven teaching and learning practices with the LEAF system includes the following:

1. Prepare the infrastructure for data-driven sustainable educational practices;
2. Support learning for all by using learning log data;
3. Foster self-directed learning skills with learning logs;
4. Enhance interpersonal communication and social learning by using learning log data; and,
5. Share data and evidence.

Our development agenda aims to overcome the challenges in realizing these tenants, which is informed by technology research as well as contributes to investigating the impact in the educational sector.

Educational Explainable AI Tools (EXAIT) is one of the services linked to LEAF.[2] EXAIT is used to learn English and mathematics at the Japanese junior high and high school levels. For English, the tool recommends appropriate extensive reading materials based on students' vocabulary exposure to prior reading materials. For mathematics, it aims to detect learners' "stuck" points based on analyzing digital pen stroke data and students' self-explanations. In addition to visualizing the learning logs, the evidence extraction function in LEAF can compute the effectiveness of interventions during daily activities as a step to promote evidence-based practices with daily log data.

Another service linked to LEAF is the GOAL system,[3] developed to provide approaches to learner modeling and support for self-directed educational activities using learners' own data synthesized from multiple data sources. LEAF also integrates cutting-edge technologies such as blockchain to facilitate validating and sharing of learning logs across institutions and utilizing logs for algorithmic grouping for cooperative learning scenarios.[5]

In Japan, the LEAF system is implemented on the national educational cloud, which is accessed by 20 public schools from seven prefectures. Additionally, instances are implemented locally in 10 universities reaching more than 20,000 students in the K–16 context for their daily learning activities.[2] Internationally, LEAF is adopted by teachers in collaborating universities in Taiwan, India, and Hong Kong[4] and enables a sustainable way to continue collecting educational Big Data in international contexts. It pursues cutting-edge research on educational data science utilizing daily educational activities and aims to support evidence-based learning and teaching practices internationally. The teachers gain access to their own class data and researchers among them also utilize the data to analyze and report findings as academic publications.

In Japan, the LEAF system is implemented on the national educational cloud, which is accessed by 20 public schools from seven prefectures.

The LEAF approach and its implementation at scale from Japan is only possible with collaboration among researchers, practitioners, and policymakers. While the GIGA school provided devices and infrastructure, LEAF provides actual applications to evolve a data-driven educational ecosystem that is unique. The research group is engaged in data collection and standardized logging policymaking. It can be adopted by industries that create online learning tools and services.

A formal council on Evidence-driven Education (EDE) was formed in 2021. EDE provides both a research forum for the different Japanese stakeholders to collaborate with LEAF and a commercial forum to adopt LEAF at the institutional and national levels through a subscription-based financial model. That large-scale adoption of the LEAF system highlights its flexibility and low threshold. It is used by teachers from various disciplines and sociocultural backgrounds. The East Asian and Oceanian community of educators and researchers can explore, adopt and have a local implementation of LEAF in different countries.[5] Our research aims to create an educational data ecosystem to continuously improve practices by involving different stakeholders who utilize data-driven services in their daily learning context.

Acknowledgment. This research work is partially supported by the following research grants: NEDO Special Innovation Program on AI and Big Data JPNP18013, JPNP20006 and JSPS KAKENHI 20H01722, 22H03902.

References

1. Flanagan, B. and Ogata, H. Learning analytics platform in higher education in Japan. *Knowledge Mgmt & E-Learning 10*, 4 (2018); 469–484.
2. Ogata, H., Majumdar, R., Flanagan, B., and Kuromiya, H. Learning analytics and evidence-based K12 education in Japan: Usage of data-driven services and mobile learning across two years. *Intern. J. Mobile Learning and Organisation* (2022); https://doi.org/10.1504/IJMLO.2023.10048714
3. Majumdar, R., Yang, Y., Li, H., Flanagan, B. and Ogata, H. 3 Years of GOAL project in public school: Leveraging learning & smartwatch logs for self-directed learning. Practitioner's Report in *Proceedings of LAK23*.
4. Ogata, H., Majumdar, R., Yang, S.J. and Warriem, J.M. Learning and evidence analytics framework (LEAF): Research and practice in international collaboration. *Info. and Tech. Education and Learning 2*, 1 (2022); https://doi.org/10.12937/itel.2.1.Inv.p001
5. Other works related to LEAF project: https://www.let.media.kyoto-u.ac.jp/en/publications-3/; explanatory video https://www.youtube.com/watch?v=LTJSMWoXH8I; to request demo account; https://sites.google.com/view/ederc/english

Hiroaki Ogata is a professor at Kyoto University, Japan.

Rwitajit Majumdar is a senior lecturer at Kyoto University, Japan.

Brendan Flanagan is an associate professor at Kyoto University, Japan.

AI in Vietnam | DOI:10.1145/3589948

Building and Nurturing AI Development in Vietnam

BY HUNG BUI, MINH HOAI NGUYEN, DAT QUOC NGUYEN, LINH PHAM, AND DINH PHUNG

IS IT POSSIBLE for a developing country like Vietnam to be a competitive player on the world stage in cutting-edge artificial intelligence (AI) research and development? Will it be able to tap into the $US15.7 trillion projected for the AI global economy by 2030? For Vietnam, these questions often went unchallenged; contemplating answers was daunting. VinAI Research, however, aims to embrace these challenges by laying the groundwork for AI innovation and growth for the region.

Founded in 2019, VinAI leapfrogged to the 20th ranking on Thundermark Capital's list of "Global AI Research Companies" by 2022, and was the only Southeast Asian (SEA) representative on the list.[a] Today, its research has been shared in more than 100 publications from top AI venues spanning across three main areas: machine learning (ICML, NeurIPS, ICLR), computer vision (CVPR, ICCV, ECCV), natural language processing (ACL, EMNLP, InterSpeech), reflecting some of the highest concentrations of AI expertise worldwide. Coincidentally (or not), Vietnam, which had never appeared on global AI research rankings before, jumped to 26th worldwide the same year.

AI recognizes facial expressions to promptly issue warnings when the driver is tired and sleepy.

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING PHOTO BY KITTISAK JIRASITTICHAI

Research at VinAI aims to push the frontiers of AI and disrupt the status quo. We question the theoretical foundations and practices in machine learning, and deep learning, and we investigate how they can enable new AI technology in natural language understanding and computer vision—two areas of applications fundamental to cognitive AI capabilities with unique challenges in our geographic locations. Our leading research exemplars in machine learning, include the young mathematics field of optimal transport for AI, self-supervised learning and domain adaptation, and in computer vision, including 3D vision, robust AI, few-shot problems, and image restoration and enhancement. In natural language processing (NLP), we are the creator of the top Vietnamese machine translation system as well as many large language models for Vietnamese.

We are also naturally drawn toward problems of developing countries, which might otherwise be overlooked in the research community. Take low-resource (LR) language models, for example. At this writing, ChatGPT has stunned the world and completely dominates social media.[b] There is no denying the success of large language models is easily one of the biggest highlights of AI research in recent years. But with close to 200 billion parameters to be trained, trillions of words needed, and millions of dollars to train a model such as GPT-3, what are the chances for LR languages like Vietnamese with little, and underpresented, data available? Our goal is to not only create new tools and knowledge agonistic to LR languages in the region, but also to create the best NLP technology for Vietnamese. As a result, we introduced our first public large-scale monolingual language model pretrained for Vietnamese in 2020.[c] Named after the famed Vietnamese noodle soup, PhoBERT[1] was built using

a http://bit.ly/3YZ2EaZ

b https://openai.com/blog/chatgpt/

c https://github.com/VinAIResearch/PhoBERT

a 20GB pretraining dataset of 145 million Vietnamese word-segmented sentences. Multiple downstream NLP tasks such as part-of-speech tagging, named entity recognition, dependency parsing, natural language inference, and text classification have achieved state-of-the-art performances based on PhoBERT. It is publicly available with 100K+ downloads per month.[d]

Another highly challenging problem for the region is language translation, which includes not only the traditional machine translation (MT, foreign text to local text), but also speech translation (S2T, foreign language speech to local text). Vietnam has achieved rapid economic growth in the last two decades, becoming an attractive destination for tourism, trade, and investment. Understandably, one common hurdle identified by both foreign and domestic business is the communication barrier—as Vietnamese (or ngôn ngữ' tiếng Việt) is also an incredibly complex language to learn.

Latin-based with six different tones, a sentence can take a totally different meaning in Vietnamese without correct tonal usage. Indeed, without proper use of tones, 'muoi', for example, can mean salt (muối), mosquito (muỗi), each (mỗi), lip (môi), or fishing bait (mồi). High-quality text and speech translation to the target language, such as Vietnamese, has become more important than ever. The current publicly available datasets for Vietnamese are nowhere near what is needed to achieve quality close to human-level translation. Once again, we are pushing the state of the art and making them available to the community: For English-to-Vietnamese, PhoMT is the first high-quality, large-scale parallel dataset for MT, consisting of 3.02M sentence pairs;[e] and PhoST consists of 508 audio hours, 331K triplets (audio, script, target language), for S2T.[f] Combined, the VinAI Translate system[2] obtains a state-of-the-art performance for each translation direction and outperforms Google Translate in both automatic and human evaluations; it is now available for public use.[g]

Vietnam, and SEA countries, are also (in)famously known for their traffic and mobility pain points. It is estimated that in 2019, 7,600 fatalities occurred on Vietnam roads,[h] with 60% a direct result of driver distractions[i] and 20% of driver drowsiness.[j] This is uniquely challenging due to mixed mode of travel, traffic conditions, local conditions, and unexpected behaviors of commuters. VinAI's research aims to create products that, quite simply, save lives. Translated from research in embedded computer vision, Driver and Occupant Monitoring (DMS)[k] technology ensures a safe driving experience with our in-cabin solutions, using low-cost cameras and highly efficient AI to analyze driver behaviors to warn against and prevent driving errors. We introduced the world's first patented Auto Mirror Adjustment (AMA) technology at CES 2023—an AI-based feature to automatically adjust the mirror to an optimal position with a single press of a button. A standard approach would need at least two cameras to triangulate, but our technology can predict eye position precisely with just one infrared camera. Beyond this, a full range of DMS includes highly accurate facial recognition for theft prevention, driver drowsiness and attention warning, advanced driver distraction warning, and dangerous behavior detection. DMS can run on multiple SoCs such as Nvidia, Qualcomm, Renesas, Ambarella, and others.[l] Another cutting-edge product provided a real-time, 360-degree view around the vehicle, thereby completely removing all blind spots not only around but underneath the vehicle. Combined with AI that can detect the presence of pedestrians, cars, trucks, motorbikes, and small children, it is a product that significantly improves safety, especially for big SUVs, trucks, and commercial buses.

Vietnam has seen significant progress in AI in recent years, as the country has made efforts to advance its technology industry and invest in R&D in the field. VinAI has also established partnerships with leading technology companies and organizations to support its AI development efforts. Additionally, the government has implemented policies to support the growth of the technology sector and attract top talent in the AI field.

In conclusion, the use of AI in Vietnam has the potential to greatly benefit the country's economy and improve various industries such as communication and transportation. However, there are also challenges that must be addressed, such as the need for adequate regulations and the need to upskill the workforce to ensure the implementation of AI is done effectively and ethically. By addressing these challenges and fully embracing the potential of AI, we are confident Vietnam can become a leader in the field and reap the rewards of this technology.

Acknowledgments. Our work at VinAI is attributed to the dedication and expertise of all the VinAI members, including, but not limited to: Anh Tran, Tung Pham, Toan Tran, Son Hua, Rang Nguyen, Toan Bui, Khoi Nguyen, Dung Nguyen, Chan Vu, Vuong Cap, Duoc Nguyen, Tuan Le, Tri Pham, Viet-Anh Nguyen (The Chinese University of Hong Kong), Cuong Pham, Thien Huu Nguyen (University of Oregon, Eugene, USA), Van Anh Dam, and Ngoc Tran.

References

1. Nguyen, D.Q and Nguyen, A.T. PhoBERT: Pre-trained language models for Vietnamese. In *Proceedings of the 2020 Findings of the Assoc. Computational Linguistics*, 1037–1042.
2. Nguyen, T.H. et al. A Vietnamese-English neural machine translation system. In *Proceedings of the 23rd Annual Conf. Intern. Speech Communication Assoc. Show and Tell*, 2022, 5543–5544.

Hung Bui is the Founder and CEO at VinAI in Hanoi, Vietnam. He is also an adjunct professor at Monash University, Clayton, Australia.

Minh Hoai Nguyen is the head of Computer Vision Group at VinAI in Hanoi, Vietnam. He is also an associate professor at Stony Brook University, NY, USA.

Dat Quoc Nguyen is Senior Research Scientist and the head of the NLP Group at VinAI, Hanoi, Vietnam.

Linh Pham is AI Residency Program Manager at VinAI, Hanoi, Vietnam.

Dinh Phung is a professor at Monash University, Clayton, Australia and has served as a senior technical consultant to VinAI, Hanoi, Vietnam, since 2021.

d Statistics available at https://huggingface.co/vinai

e https://github.com/VinAIResearch/PhoMT

f https://github.com/VinAIResearch/PhoST

g https://vinai-translate.vinai.io/

h http://bit.ly/3JduYR0

i http://bit.ly/3LjaUz4

j https://bit.ly/3yyWMdp

k https://www.youtube.com/watch?v=qY7c1e3lo7s

l http://bit.ly/3l830hm

Feature Computing | DOI:10.1145/3589224

Principles and Practices of Real-Time Feature Computing Platforms for ML

BY HAO ZHANG, JUN YANG, CHENG CHEN, SIQI WANG, JIASHU LI, AND MIAN LU

REAL-TIME FEATURE COMPUTATION, which calculates features from raw data on demand, is a crucial component in the machine learning (ML) application process. These real-time features are vital for various real-world ML applications, such as anti-fraud management, risk control, and personalized recommendations. In these cases, low latency (milliseconds) in computing fresh data features is crucial for accurate and high-quality online inference.

As illustrated in the accompanying figure, a data scientist typically begins an ML application by developing feature computation scripts (for example, using Python or SparkSQL) for offline training. However, these scripts cannot meet the demands of online serving, including low latency, high throughput, and high availability. Hence, it is necessary to transform these scripts into performance-optimized code (for example, using C++) that can be developed by an engineering team with system and production knowledge. This transformation process is time-consuming and requires significant development, deployment, and double system maintenance efforts.

Moreover, the big gap in the software stacks, personnel domain knowledge, and performance concerns may cause the challenging feature inconsistency problem; for example, different interpretations of window boundary (that is, inclusive or exclusive), variations in handling empty values, or diverse operator definitions. As an example, consider fraud detection where the features include "account balance" and its standard deviation (std). During offline training, the model may use yesterday's account balance while in online serving, the model is provided with the current account balance.[1] Moreover, the std logic may also differ, as shown in the figure. These variations in data definition and computing can result in significant differences in prediction outcomes (that is, training-serving skew). Consequently, data scientists and engineers must invest significant effort in iterative development and cooperative consistency verification to align the results. To resolve these challenges, new design methodologies and systems are necessary to eliminate the added cost of consistency verification and ensure low latency in on-demand feature computing.

OpenMLDB relieves the headache of verification by offering a unified execution engine and the same SQL APIs for both offline training and online serving.

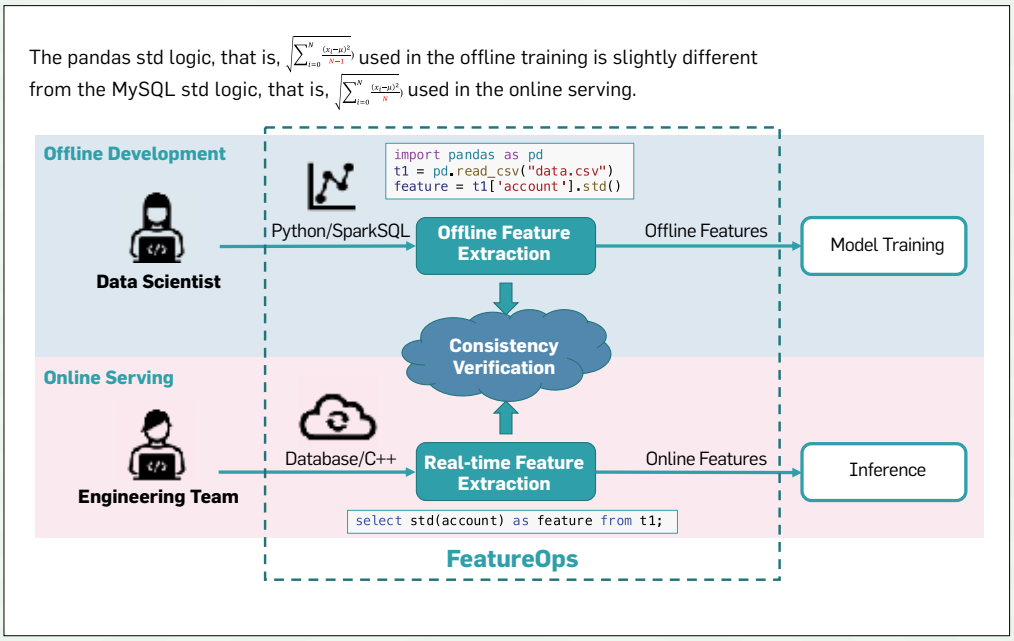
In this article, we discuss the design principles and practices of tackling the challenges of real-time feature computing. Basically, we propose that such a feature computing platform should consist of three major components—a batch engine for offline training; a real-time engine for online serving; and a unified execution plan generator—to bridge those two engines to inherently guar-

IMAGE BY CKA/SHUTTERSTOCK

antee the online-offline consistency. With such an architecture, a feature script developed by a data scientist can be deployed online without extra effort involved.

OpenMLDB[a] is an open source feature computing platform practicing those design principles, which is initiated and led by Fourth Paradigm Southeast Asia. In particular, the unified SQL APIs and shared execution plan generator for both offline and online engines eliminate the transformation and consistency verification.

With optimization techniques such as in-memory time-travel structures, LLVM codegen, and pre-aggregation, the real-time engine of OpenMLDB offers low latency and can meet the demands of most online decision-making systems with an average latency of under 20ms. However, this is not achievable by other feature stores (for example, Feast,[b] Hopsworks,[c] Feathr[d]) or traditional databases (for example, MySQL), which either provides only storage service for fast retrieval or cannot satisfy the latency requirement of on-demand feature computing. By taking advantage of the Persistent Memory Module (PMEM), OpenMLDB can further shorten the tail latency up to 19.7%, reduce the recovery time up to 99.7%, and save up to 58.4% of the total cost compared to the version of DRAM+SSD.[2] Moreover, OpenMLDB is considered for state-of-the-art federated learning systems.[3] It has been integrated into Fourth Paradigm's commercial products used by its customers (for example, Epitex), and also independently used by an increasing number of community users (for example, Akulaku) since it is open source.



Consistency verification between offline and online feature computing.

Here, we examine a real-world use case from Akulaku (a fintech unicorn from Indonesia) to demonstrate how OpenMLDB helps the company's products. The biggest challenge faced by Akulaku in its ML deployments is the consistency verification before new features are employed online. Debugging the cause of a training-serving skew was a time-consuming and labor-intensive process, taking up to a month of manpower and 50% of their total workload. Moreover, verification failures could lead to ineffective models or even serious production accidents. OpenMLDB relieves the headache of verification by offering a unified execution engine and the same SQL APIs for both offline training and online serving. It helps optimize their architecture by eliminating the various external tools (for example, Spark, Greenplum for the offline features and TiDB, PolarDB, Flink for the online features) in their original software stack. Reported by Akulaku, OpenMLDB helps them save more than $US500K per year in terms of server and personnel costs.

In addition to the users in Southeast Asia, OpenMLDB has also been adopted by users from other regions, such as Huawei, 37GAMES, UBiX, and so on. As feature engineering becomes a crucial component in ML applications, we believe that OpenMLDB will play a significant role in accelerating ML deployment and productization not only in Southeast Asia but worldwide. C

OpenMLDB helped Akulaku save more than $500,000 per year in server and personnel costs.

a https://github.com/4paradigm/OpenMLDB
b https://feast.dev/
c https://www.hopsworks.ai/
d https://github.com/feathr-ai/feathr

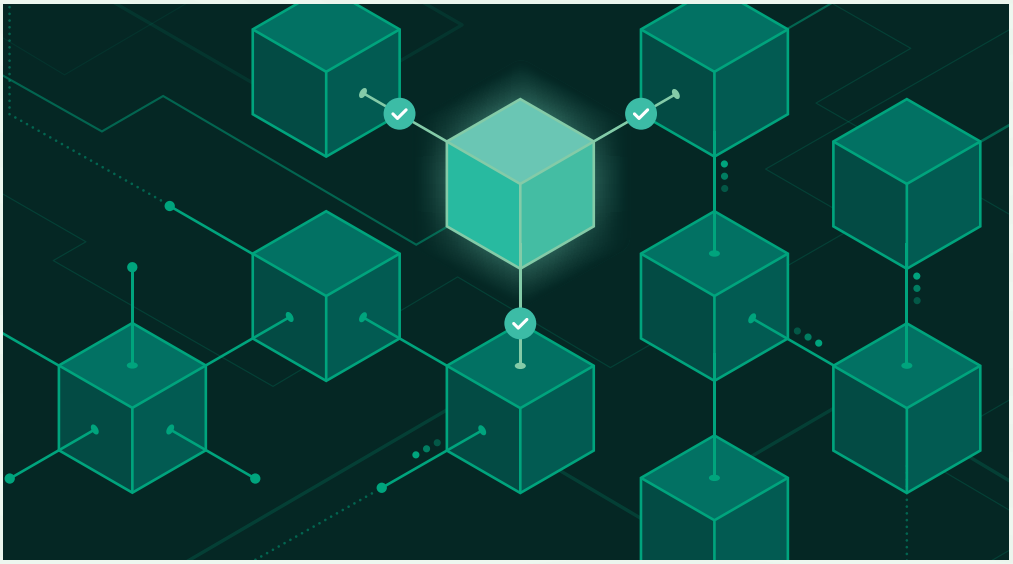
References
1. Agrawal, R. and Lee, B. Feature store: Challenges and considerations; http://bit.ly/3YHkQ8R.
2. Chen, C. et al. Optimizing in-memory database engine for AI-powered on-line decision augmentation using persistent memory. In *Proceedings of the VLDB Endow. 14*, 5 (Jan. 2021), 799–812.
3. Wang, S., Li, J., Lu, M., Zheng, Z., Chen, Y. and He, B. A system for time series feature extraction in federated learning. In *Proceedings of the 31st ACM Intern. Conf. Information & Knowledge Management*, 2022.

Hao Zhang, Jun Yang, Cheng Chen, Siqi Wang, Jiashu Li, Mian Lu, Fourth Paradigm Southeast Asia Pte. Ltd., Singapore.

Blockchain Protocols | DOI:10.1145/3589225

The Future of Blockchain Consensus

BY VINCENT GRAMOLI AND QIANG TANG



THE BYZANTINE CONSENSUS problem was originally defined in 1982. The idea is for correct machines (or nodes) to propose values and reach a consensus despite the presence of malicious (or Byzantine) ones. More precisely, three properties must be satisfied: these correct nodes must eventually decide (termination); their decision should be identical (agreement); and the value they decide should be one of the proposed values (validity). Since then, there have been many protocols for different models.

Most of the classical consensus solutions were unfortunately designed for local area networks and were typically illustrated with four well-connected nodes running a network file system. With the advent of blockchain technologies, the consensus problem became crucial for blockchain nodes to agree on the same blocks. But blockchains are deployed among many nodes spread across the open Internet, which naturally raises new challenges like making consensus protocols robust and efficient at large scale and in non-purely synchronous networks. As the classic consensus protocols were not designed for the qualitatively different large and open Internet, the blockchain consensus faced limitations and required new perspectives to address them.

Our recent work has led to significant advancements that push blockchain consensus to be truly scalable and robust, while preserving high performance.

The Redbelly path to making blockchain consensus truly scalable. The crux of the problem that prevented these recent blockchain systems from performing efficiently at large scale lies in the fact that these solutions typically solve the traditional validity guarantee from 1982, hence forcing blockchain nodes to decide a single block at each index, regardless of the number of proposed blocks. It is thus extremely difficult for the blockchain performance to increase with the amount of node resources.

Our new blockchain protocols emerged with the intent to solve consensus and to scale to large networks. The idea, illustrated by the Redbelly Blockchain[2] from the Commonwealth Science Industry and Research Organisation (CSIRO) and The University of Sydney, consists of solving a slightly different variant of the consensus problem, deciding a superblock composed of multiple proposed blocks. This led to significant scalability improvements as the number of decided transactions could finally scale linearly with the number of participating nodes. As a result, the Redbelly Blockchain demonstrated that solving consensus did not prevent scaling to 1,000 machines geo-distributed worldwide.

The Australian federal government included the Redbelly Blockchain as part of its National Blockchain Roadmap. To strengthen security, the consensus protocol of Redbelly Blockchain was formally verified with model checking,[1]

The Redbelly Blockchain demonstrated that solving consensus did not prevent scaling to 1,000 machines geo-distributed worldwide.

IMAGE BY ELENABSL/SHUTTERSTOCK

which drastically reduces the risks of human errors that have been apparent in other blockchains. Some recent results demonstrate that, while modern blockchains could not execute real traces on decentralized applications,[a] Redbelly Blockchain could.[5] The techniques to make blockchains scalable and solve consensus are now part of a book and taught online to thousands of students in the University of Sydney's blockchain scalability MOOC on Coursera. More details can be found at https://gramoli.github.io.

The Dumbo path to make asynchronous consensus real. Most of the consensus in use can guarantee to work only in synchronous networks (assuming all messages to be delivered within some time parameter). This makes them vulnerable in the open Internet, which is naturally asynchronous because of dynamic network conditions or even network attacks. The reason asynchronous consensus is not used yet is due to its complexity, as hinted by the famous FLP-impossibility that deterministic asynchronous consensus does not exist. Over the years, a great deal of effort has been devoted to design various randomized protocols to circumvent the impossibility. Unfortunately, most of them are theoretical in nature and not implemented, until the recent HoneyBadgerBFT, which essentially asked the community whether asynchronous consensus can ever be practical.

Recent progress of the Dumbo protocol family affirmatively answers the question. Starting with DumboBFT,[3] major bottlenecks of HoneyBadgerBFT (that makes it much slower when scales up) were identified, and a new design principle was proposed to use a different tool called multi-valued validated Byzantine agreement (MVBA) to conquer the obstacle to achieve only a small constant number of rounds (no matter the scale). In Dumbo-MVBA,[b] a long-standing open problem of communication optimal MVBA was finally resolved. More importantly, the new design methodology further enables a more compact and even faster Speeding Dumbo,[c] and a tight broadcast-and-vote structure for asynchronous consensus to realize high throughput without sacrificing latency in Dumbo-NG,[d] which harvests most of the bandwidth.

Our new blockchain protocols emerged with the intent to solve Internet consensus, that is, to scale to large networks and to be robust in the asynchronous network while preserving high performance.

Considering deterministic synchronous consensus protocols are very fast (but may fail to work in the open Internet), while asynchronous consensus is robust (slower), the recent combination brought out the best of both. Specifically, a new practical framework for optimistic asynchronous consensus called Bolt-Dumbo-Transformer[4] was proposed. When deployed and tested over 100 nodes spread across five continents, Dumbo protocols show orders of magnitude advantages over previous protocols and clear evidence for practical performance. Wide media coverage includes CoinDesk and multiple major media outlets covering Dumbo protocols. More details can be found at https://alkistang.github.io. C

a https://diablobench.github.io/

b https://eprint.iacr.org/2020/842

c https://eprint.iacr.org/2022/027

d https://arxiv.org/abs/2209.00750

The Redbelly Blockchain research team at the University of Sydney.

References

1. Bertrand, B., Gramoli, V., Lazić, M., Konnov, I., Tholoniat, P., and Widder, J. BA: Holistic verification of blockchain consensus. In *Proceedings of the 2022 ACM Symp. Distributed Computing*.
2. Crain, T., Natoli, C., and Gramoli, V. Red Belly: A secure, fair and scalable open blockchain. *IEEE Symposium on Security and Privacy* (2021), 466–483.
3. Guo, B., Lu, Z., Tang, Q., Xu, J., and Zhang, Z. Dumbo: Faster asynchronous BFT protocols. In *Proceedings of the 2020 ACM Conf. Computer and Communications Security*.
4. Lu, Y., Lu, Z., and Tang, Q. Bolt-Dumbo transformer: Asynchronous consensus as fast as the pipelined BFT. In *Proceedings of the 2022 ACM Conf. Computer and Communications Security*.
5. Tennakoon, D., Hua, Y., and Gramoli, V. Smart Red Belly blockchain: Reducing congestion for Web3. In *Proceedings of the 37th IEEE Intern. Parallel & Distributed Processing Symp.*, 2023.

Vincent Gramoli is an ARC Future Fellow at the University of Sydney and the founder and CTO of Redbelly Network.

Qiang Tang is a Senior Lecturer and SOAR Fellow at the University of Sydney, Australia.

Cyber-Physical Systems | DOI:10.1145/3589648

Challenges in Designing Blockchain for Cyber-Physical Systems

BY GUNTUR DHARMA PUTRA, SIDRA MALIK, VOLKAN DEDEOGLU, RAJA JURDA, AND SALIL S. KANHERE

With a population of 2.37 billion[a] and a GDP of more than $US30 trillion,[b] East Asia and Pacific's economic growth has seen a massive increase in cooperation and competition in recent years. To foster more growth in the region, there must be an automation of business processes that guarantees fair and open collaboration across multiple stakeholders. Cyber-physical systems (CPS) can enable automation across cyber and physical domains. In addition, blockchain technology offers auditability, transparency, and decentralization, thus showing promise to enable further developments in the region's economy. Combining these technologies would allow for a broad range of unprecedented possibilities. However, incorporating blockchains for CPS is challenging.

In the last couple of years, we have conducted cutting-edge research across Australian institutions—namely UNSW Sydney, QUT, and CSIRO's Data61—where we explored addressing these challenges in different verticals (domains) and horizontals (challenges), as seen in the accompanying figure. In this article, we discuss some of our work to share our experience in tackling the specific challenges of scalability, trust, and privacy.

a https://bit.ly/3B9Iguf
b https://bit.ly/42Ithmq

In PrivChain, blockchain plays an instrumental role in verifying proofs, initiating relevant payments, and logging the results on-chain.

Scalability

The append-only nature of blockchain raises concerns about scalability. With the passage of time, more records will be appended to the blockchain, which would consume a considerable amount of storage. Complex CPS applications, such as supply chain management, involve many participants requiring rapid read-write operations, necessitating a scalable blockchain.

In ProductChain,[3] we demonstrated how we addressed the scalability issues in blockchain-supported supply-chain management by leveraging parallel chains instead of a single large chain combined with a tiered architecture for access management. The idea of having parallel chains, called shards, is borrowed from parallel processing, where each shard maintains its own synchronized ledger, called a local ledger. ProductChain organizes the shards based on geographic zones, which are derived from the locality of the trade. To obtain a consolidated view, a global validator is responsible for returning on-chain data across multiple shards.

Trust

While blockchain enables trust-less interaction, it does not guarantee the trustworthiness of data, which may contain noise or be manipulated by a malicious entity, diminishing its veracity. The immutability of blockchain inherently exacerbates this condition, as inaccurate on-chain data cannot later be corrected. CPS applications

IMAGE BY ICO MAKER

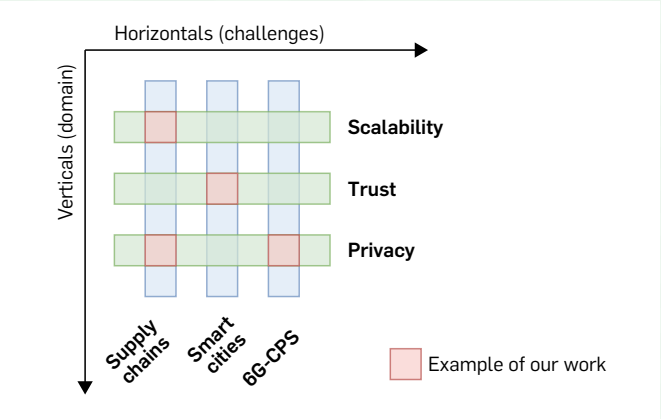


Diagram illustrating the extent of the challenges in incorporating blockchain for CPS applications, which spans across various verticals (blockchain scenarios or domains) and horizontals (the inherent challenges).

thus require holistic trust mechanisms to ensure the end-to-end integrity of the collected data and the associated interactions.

To tackle this issue, we designed a layered trust architecture for generic blockchain-enabled CPS applications.[1] We demonstrated how to afford an end-to-end trust mechanism by devising a reputation system for participating nodes. The architecture quantifies the trustworthiness of sensor observations by utilizing a node's confidence in its data and corroborating data from neighboring sensors. Additionally, the architecture establishes trust at the block-generation level through an adaptive block validation mechanism, increasing the overall throughput.

Privacy

Blockchain stores information on transparent chained blocks replicated across all blockchain participants. While replication improves transparency, there is a trade-off between transparency and data protection. In the case of private blockchains, the information stored on-chain is sometimes associated with the identity of the contributing stakeholders to ensure trust. However, identity association makes it difficult to protect sensitive information. In the case of public blockchains, using pseudonyms introduces a level of anonymity, concealing the real identity from the public. However, blockchain de-anonymization is still possible by linking transactions with public keys used consecutively.

One solution is to generate and share computations and proofs of the data rather than the sensitive data itself. With PrivChain,[2] a privacy-preserving blockchain-based supply chain management system, we demonstrated how to achieve that scenario. PrivChain leverages zero knowledge proofs (ZKP) and commitment schemes to resolve the privacy trade-off issue. ZKP allows someone to prove that some secret information is accurate without disclosing any detail about the secret, while commitment schemes allow for verification of a committed secret later in time. Supply-chain participants can thus share the proof of their supply-chain data related to their products instead of the raw data, which may pose privacy risks. In PrivChain, blockchain plays an instrumental role in verifying these proofs, initiating relevant payments, and logging the results on-chain. The verification of the proofs is automated on-chain through smart contracts, self-executing deterministic computer programs that run as per predetermined rules and logic.

Another solution would be to use different pseudonyms for each transaction in a public blockchain. We demonstrated how to achieve a privacy-preserving reputation system for a 6G-CPS,[4] where each node utilizes changeable public keys in each transaction. To provide stronger privacy preservation, our reputation system utilizes interconnected public-private blockchains, using which nodes may store sensitive information on the private chain, while publicly viewable information is stored on the public chain for auditability purposes.

Research Impact and Future Directions

With our work addressing the issues mentioned here, the region could increase the benefits from blockchain and CPS in realizing fair competition with secure and transparent cross-border transactions. For instance, food producers from various locations in East Asia and Pacific could use ProductChain[3] to facilitate open and secure trade operations, removing unnecessary paperwork that hinders efficient collaborations. Our reputation system[1] can also be implemented as an additional layer of security, where reputation scores indicate the trustworthiness of the food producers in the supply chain network.

However, there remain abundant research opportunities for further exploration. Future work may investigate more efficient techniques for addressing these inherent challenges. For instance, to tackle unresolved scalability issues, further research may explore Layer-2 blockchain solutions, that is, efficiency and trustworthiness of computing operations that occur off-chain. It would also be worthwhile to explore blockchain-based data-sharing solutions for multiple application domains within CPS. C

References

1. Dedeoglu, V., Jurdak, R., Putra, G.D., Dorri, A. and Kanhere, S.S. A trust architecture for blockchain in IoT. In *Proceedings of the 16th EAI Intern. Conf. Mobile and Ubiquitous Systems: Computing, Networking and Services*, (Houston, TX, USA, Nov. 2019). ACM, 190–199; doi: 10.1145/3360774.3360822.4
2. Malik, S., Dedeoglu, V., Kanhere, S.S. and Jurdak, R, PrivChain: Provenance and privacy preservation in blockchain enabled supply chains. In *Proceedings of the 2022 IEEE Intern. Conf. Blockchain*, Aug. 2022, 157–166; doi: 10.1109/Blockchain55522.2022.00030.
3. Malik, S., Kanhere, S., and Jurdak, R. ProductChain: Scalable blockchain framework to support provenance in supply chains. In *Proceedings of the IEEE 17th Intern. Symp. Network Computing and Applications*. (Cambridge, MA, USA, Nov. 2018), 1–10; doi: 10.1109/NCA.2018.8548322.
4. Putra, G.D., Dedeoglu, V., Kanhere, S.S., and Jurdak, R. Toward blockchain-based trust and reputation management for trustworthy 6G networks. *IEEE Network 36*, 4 (Jul/Aug 2022), 112–119; doi: 10.1109/MNET.011.2100746.

Guntur Dharma Putra is a lecturer in the Department of Electrical and Information Engineering at UGM, Indonesia. This work was done while he was a Ph.D. student at UNSW Sydney, Australia.

Sidra Malik is a postdoctoral fellow at CSIRO's Data61, Sydney, Australia

Volkan Dedeoglu is a research scientist in the Distributed Sensing Systems Group at CSIRO's Data61, Sydney, Australia

Raja Jurdak is a professor of distributed systems and chair of Applied Data Sciences, as well as Director of the Trusted Networks Lab at Queensland University of Technology, Brisbane, Australia.

Salil S. Kanhere is a professor in the School of Computer Science and Engineering at UNSW Sydney, Australia.

Visual Query Interfaces | DOI:10.1145/3589153

To Draw Is Human: Toward No-Code Subgraph Search

BY SOURAV S. BHOWMICK AND BYRON CHOI

DUE TO THE worldwide shortage of developers, growing talent gap, and budgetary challenges faced by small- and medium-sized businesses in hiring software teams, low-code or no-code frameworks are the latest disruption in the business world.[1] For example, SAP recently launched SAP AppGyver, which is a "no-code application development platform that enables developers of all skill levels to create enterprise-ready applications with drag-and-drop simplicity."[5]

The demand for such low-code or no-code frameworks is not limited to software applications development but also for easy access and search of data residing in databases. Specifically, lay users should be able to access them without needing to write a single line of code. However, query languages (QL)—the primary means to access data residing in databases—enforce end users to be proficient in these languages before they can take advantage of databases for their tasks.

Visual query interfaces (VQIs) are popular low-code/no-code frameworks that facilitate access and search of databases by *drawing* queries with "drag-and-drop simplicity" instead of writing them using a QL. Given the ubiquity of graphs to model data in a wide variety of domains (for example, biology, chemistry, ecology, social science, and journalism), we summarize our 15-year odyssey that began in 2008, long before low-code/no-code frameworks became an industry disruption, to address unique research challenges brought by such frameworks in subgraph search environment. Specifically, subgraph search query is one of the most popular query paradigms for accessing graph data. Since graphs are intuitive to draw, graph data management tools from academia and industry (for example, PubChem[a]) are increasingly exposing VQIs to enable an end user to draw a subgraph search query interactively without writing a single line of code in a proprietary QL.

Visual query interfaces are popular low-code/no-code frameworks that facilitate access and search of databases by *drawing* queries with "drag-and-drop simplicity" instead of writing them.

We focus on two novel streams of research that exploit the central role played by no-code visual environments in subgraph search. First, classical VQI frameworks for low-code/no-code search suffers from several limitations, such as high creation and maintenance cost, lack of superior support for no-code subgraph query formulation, and poor portability across application domains and data sources.[2] We introduce the paradigm

a PubChem Sketcher; http://bit.ly/3mYpIt6

IMAGES BY DMITRY KOVALCHUK

of plug-and-play (PnP) VQI that addresses these limitations by automatically constructing a no-code VQI framework for a graph repository.[2] It departs from the traditional mantra of "manual" VQI construction using software developers by automatically generating and maintaining a VQI for a given graph data source in a data-driven manner without resorting to any coding. A user can simply "plug" a PnP interface on top of their graph data source (that is, socket) and "play" by drawing subgraph queries using drag-and-drop. In particular, the design of PnP interfaces is grounded on well-founded principles of human-computer interaction (HCI) and cognitive psychology to enhance the usability and reach of no-code subgraph querying frameworks.

Second, given a classical or PnP VQI that supports no-code visual querying, we explore the unique opportunity of *blending* visual query formulation with query processing.[3] Specifically, we interleave (that is, blend) query construction and query processing to prune false results and prefetch partial query results by exploiting the latency offered by the drag-and-drop activities during query formulation. The key benefits of this paradigm are at least two-fold. First, it significantly improves the system response time (SRT), which is the time a user waits to view the query results. This is because the query processor does not remain idle during visual query formulation by processing the potential subgraph query early based on "hints" received from the user. In traditional query processing paradigm, SRT is identical to the time taken to evaluate the entire query. In contrast, in this paradigm SRT is the time taken to process a part of the query yet to be evaluated (if any). Second, as a visual query is iteratively processed during query formulation, it paves the way for realizing efficient techniques that can enhance usability of graph databases, such as query suggestion, empty result feedback, and exploratory search.

Both of these research directions depart from the long-standing classical paradigm of query formulation and processing in a no-code/low-code environment. Since the inception of VQIs, they have been constructed manually by programmers. We depart from this tradition and demonstrate how it can be constructed in a data-driven manner paving the way for no-code generation of no-code subgraph search frameworks. Second, since the inception of database technology several decades ago, the classical paradigm of querying has always been served "neat." A query is formulated by an end user or application and the query processor is responsible for evaluating it once it is completely formulated. That is, query formulation precedes query processing. They were not blended. Our research opened the possibility of blending them together in the context of graph data by exploiting the unique characteristics of no-code querying frameworks, bringing in significant benefits to usability and query performance. Subsequently, some visual data systems have leveraged this paradigm of blending for data processing and analytics. For instance, Northstar[4] exploits "think time" available in an interactive VQI for approximate query processing.

The design of PnP interfaces is grounded on well-founded principles of human-computer interaction and cognitive psychology to enhance the usability and reach of no-code subgraph querying frameworks.

In conclusion, we note the two key enablers of this research—HCI and data management—have evolved into two disparate and vibrant scientific fields, rarely making any systematic effort to leverage techniques and principles from each other toward superior realization of these efforts. Through our research, we continuously strive to bridge this chasm to realize superlative and user-friendly no-code querying frameworks for diverse end users.

References

1. Atkins, B. The most disruptive trend of 2021: No code/low code. *Forbes*; http://bit.ly/3ljIxG.
2. Bhowmick, S.S., Choi, B. Data-driven visual query interfaces for graphs: Past, present, and (near) future. In *Proceedings of Intern. Conf. Management of Data*. ACM (June 2022).
3. Bhowmick, S.S., Choi, B., and Li, C. Graph querying meets HCI: State of the art and future directions. In *Proceedings of Intern. Conf. Management of Data*. ACM (June 2017).
4. Kraska, T. Northstar: An interactive data science system. In *Proceedings of VLDB Endowment 11*, 12 (2018).
5. Sindhu, B. Create, automate, and scale with low-code/no-code innovations. *SAP News Center*; http://bit.ly/429NXVj.

Sourav S. Bhowmick is an associate professor in the School of Computer Science and Engineering at Nanyang Technological University (NTU), Singapore.

Byron Choi is a professor at Hong Kong Baptist University, Hong Kong SAR, China.

Environmental Data | DOI:10.1145/3589233

DIAS—Earth Environment Data Integration and Analysis System

BY EIJI IKOMA AND MASARU KITSUREGAWA

OUR GROUP HAS been developing and operating a platform to acquire, archive, and manage various data related to the Earth's environment to make it available to researchers across a wide range of fields.

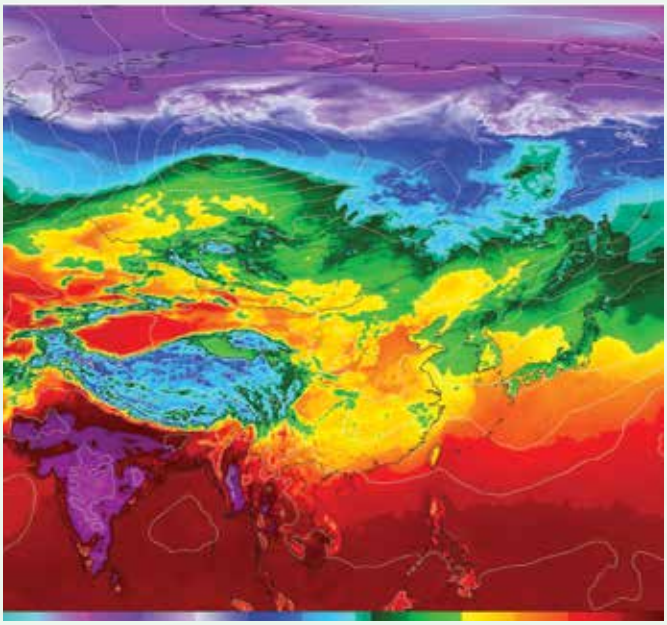
Development of this system began in the 1980s to receive, archive, and distribute Asian satellite image data. Currently, the system covers a variety of data including weather, climate change, disaster prevention, biodiversity, health, and agriculture. Today, the Data Integration and Analysis System (DIAS[a]) is a large-scale analysis platform with huge storage and more than 10,000 registered users (half of them in Japan and the other half primarily in Asia). As shown in the accompanying figure, users can easily use data collected by the common collection API through the common use API, and they can operate services at the application layer.

Archived real-time data includes observation data from automatic observation equipment, weather radars, live cameras, and forecast data generated on occasion. In many cases, once an acquisition fails, it cannot be completed. Therefore, we developed a robust tool to ensure data acquisition within a specified timeframe being used for river observation data, live camera data, and radar data. Social sensing data is another type of real-time data, for which we use a common platform developed for biodiversity research and disaster information collection applications. Non-real-time data includes observational, experimental, and simulation data generated by large-scale forecasting experiments. For efficient data uploading, we provide our own quality control and metadata input tools, which have achieved rapid data release in many international research projects, significantly shortening the time from data acquisition to data release.

The data download system, which is the most frequently used function, has been improved to add a function to extract only the necessary elements in consideration of the recent increase in data volume, and an API is also provided for real-time users. In addition, we offer an analysis environment (virtual server environment) that can directly access real-time data and very large amounts of data that is difficult to download. It allows users to login to the DIAS system and directly develop and operate services on it. This service not only reduces the enormous download time but also enables real-time analysis in an environment where data is acquired in real time.

Currently, similar data infrastructures exist, such as Copernicus in Europe and Open Data Cube in Australia. They may lack the ability to integrate different types of data or be specialized systems for specific types of data. Neither of the two systems can provide integrated data use and services as flexibly as DIAS.

Several applications that use this infrastructure are currently in operation. For example, a service that predicts river flooding and inundation by integrating and analyzing various types of real-time data archived by the DIAS and local observation data is in operation in the Philippines[b] and Sri Lanka,[c] where floods have occurred frequently in recent years. It is being expanded to Vietnam, Indonesia, Myanmar, Iran, and



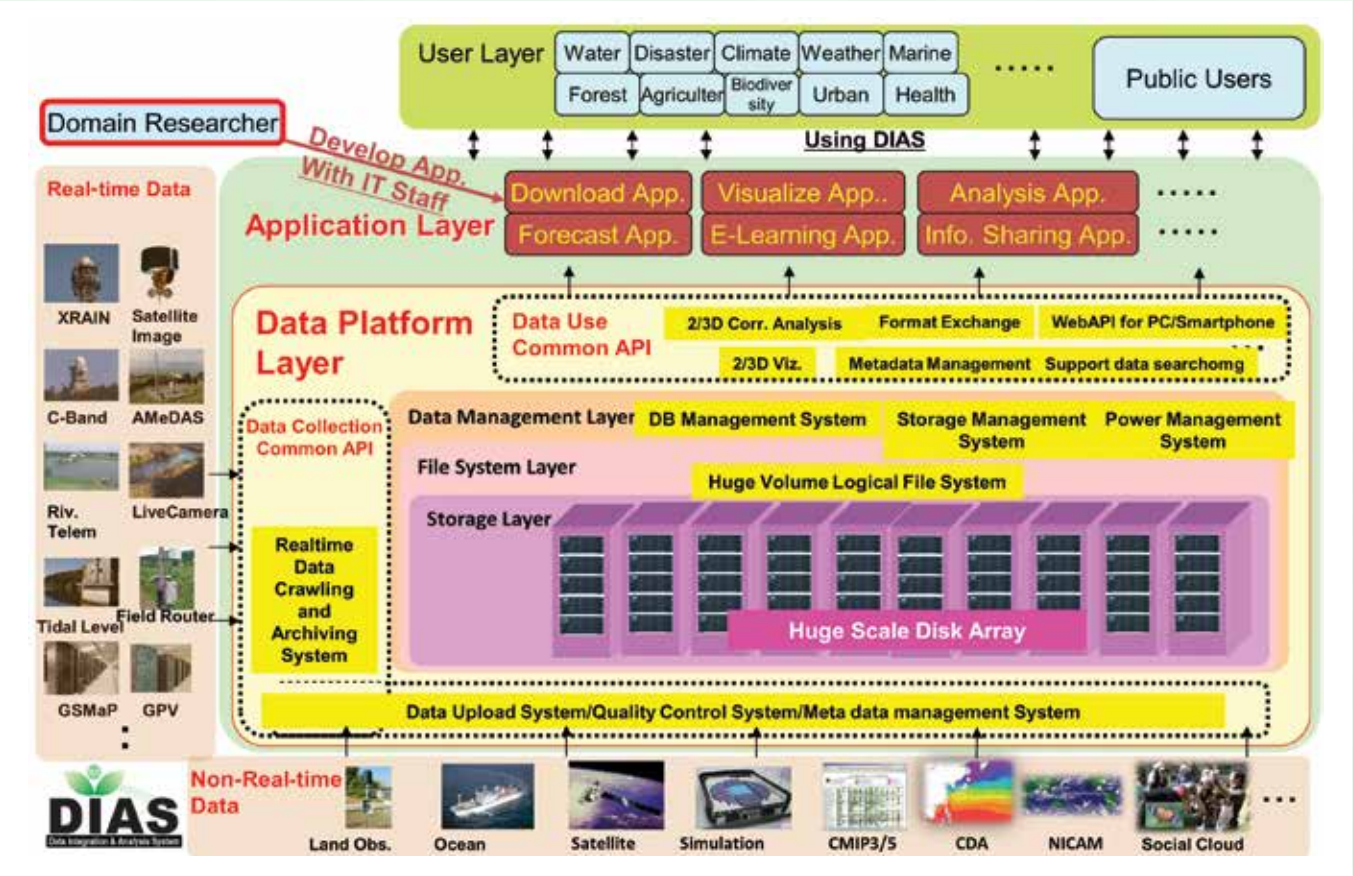
DIAS is a large-scale analysis platform with huge storage and more than 10,000 registered users.

a https://diasjp.net/en/

b https://www.mdpi.com/2073-4441/13/9/1218

c https://www.mdpi.com/2073-4441/14/6/978

IMAGE BY STOCK LPA



DIAS system structure.

the Niger-Volta River basin in Africa, where similar damage has occurred. In addition to issuing forecast information, the system is also equipped with an e-learning system for policy-makers of local institutions to learn how to operate the system, thereby contributing to human resource development in each country. In addition, the database on butterflies developed using DIAS's citizen-participatory data collection platform has been used to contribute to research on biodiversity in Japan.[d]

Because this platform is species-independent, the system is now being used for many other organisms, such as fish, shellfish, dragonflies, and storks, contributing significantly to the fields where databases have not been adequately developed. S-uiPS,[e] a real-time urban inundation forecasting application that utilizes minute-by-minute rainfall data, rainfall forecast data, and vast amounts of city block data, also uses this platform. It is now available for Tokyo[f] in response to urban disasters associated with heavy rainfall, which are particularly severe in the Asian region. This service had used a program developed by a researcher in social infrastructure engineering to calculate road flooding depths, but it took approximately 10 hours to make a 30-minute prediction. Our development team, consisting of experts in information engineering, assisted in re-writing the program to take advantage of GPUs, which resulted in an improvement of efficiency by 60 times, enabling real-time operation. This is a good example of cooperation between non-IT and IT experts and is a result of interdisciplinary collaboration. Because this service does not use region-specific parameters, it can be easily deployed in other Asian cities with problems like inland flooding, as long as the city's infrastructure information is available.

Thus, we not only provide and expand the data on DIAS but also collaborate with researchers in various fields to provide an efficient usage environment and cooperate to expand the already obtained solutions to more countries. In the future, we plan to respond to requests from data providers and support the analysis of user trends by employing access logs and other data. We will continue to expand and improve the functions of DIAS to support the research community in Asia that uses Earth environmental data. Particularly, we are analyzing the interrelationships among the data fields used, the relationship between the occurrence of sudden social events and high access data, and the design of efficient interfaces based on typical access patterns. C

d https://esj-journals.onlinelibrary.wiley.com/doi/10.1111/1440-1703.12161

e https://www.jstage.jst.go.jp/article/jscejhe/75/2/75_I_1315/_article/-char/ja/

f https://diasjp.net/service/s-uips/

Eiji Ikoma is Project Associate Professor for the Earth Observation Data Integration and Fusion Research Initiative (EDITORIA) at the University of Tokyo.

Masaru Kitsuregawa is Project Professor for the Earth Observation Data Integration and Fusion Research Initiative (EDITORIA) at the University of Tokyo.

Digital Manufacturing | DOI:10.1145/3589662

Digital Twins and Dependency/Constraint-Aware AI for Digital Manufacturing

BY DIMITRIOS GEORGAKOPOULOS AND PREM PRAKASH JAYARAMAN

INCREASING PRODUCTIVITY IN manufacturing has been an elusive goal despite significant advances in factory automation technology and robotics. There are four main challenges currently facing manufacturers: low production efficiency; product defects and inconsistent quality; unforeseen machine maintenance; and high energy use and waste costs. The fourth industrial revolution—also referred to as Industry 4.0—sets out critical technological directions for addressing these grand challenges via data-driven digital manufacturing (DM) solutions incorporating novel computing technology that combines AI/machine learning (ML) and digital twins (DTs)[4] for digitally representing complex physical industrial machine, products, and people in production.

Although the Industry 4.0 vision and directions are supported by major manufacturing companies and technology providers (for example, Siemens, Bosch, and IBM), its technology baseline is not mature enough to address related computing needs. DTs are still in the early stages of development supporting only limited cases of machine monitoring/visualization and process design/testing. Similarly, existing ML-based solutions for predicting and fixing production problems typically suffer from high inaccuracy, which is due to the complex data dependencies and constraints imposed by the process and the machines, products, and people involved, and also insufficient training data bound by achievable production runs.

While digital manufacturing powered by digital twins and dependency/constraint-aware ML is still in early stages, it has shown its potential in improving manufacturing productivity by 20%–30%.

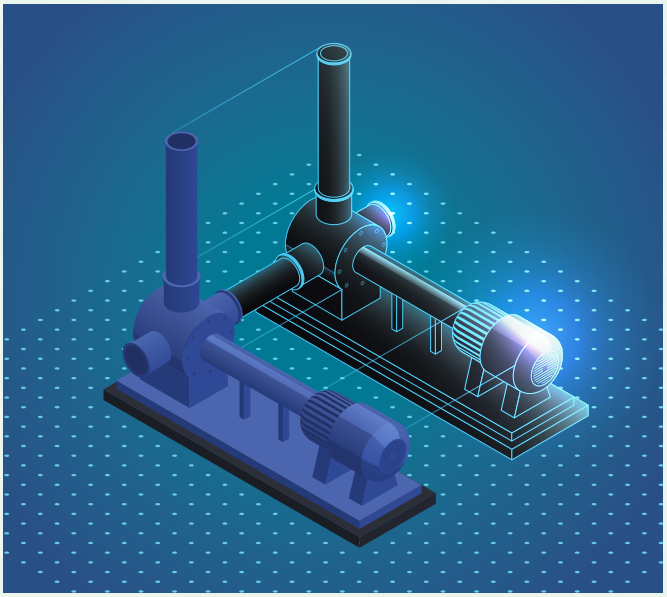


IMAGE BY MACROVECTOR

We have been developing, testing, and deploying DM solutions in food, steel, and composite manufacturing plants that incorporate the following novel research outcome in computing technologies for DM:

► **Novel DTs** for DM that digitally represent the main physical entities in manufacturing. Each DT includes a physical twin, which a complex industrial machine, a product, or a person in production; a virtual twin involves knowledge graph-based representation of the physical entity including its parts, input/output and other variables, as well as related relationships, dependencies, and constraints between them; and digital threads, which are communication channels between the two. DTs provide cyber-physical representation of the physical entities in manufacturing, support their integration in DM solutions, and enable interactions with the physical entities involved. Novel research outcomes include the knowledge graph-based models of the manufacturing entities, model-based creation of DTs, and the management of the DT life cycle.[1,2]

► **Novel dependency/constraint-aware ML (dcML)** models for DM that predict production outcomes (for example, the end-to-end production

efficiency, the quality/consistency of the product(s), maintenance issues, and the energy consumed). dcML-based predictions are used to make improvements/corrections to the production processes and related manufacturing entities (for example, by changing machine settings or reducing the speed of production) to prevent related issues before they occur. For example, an automation dependency exists between an industrial oven's temperature settings variable and the temperature variables of its temperature sensors (which are referred to as machine parameters in DM). Increasing the oven's temperature will cause the values of its temperature parameters to increase accordingly. A sample physical constraint of the oven's parameters involves the time it takes from setting the temperature to reach it. Novel dcML models are designed to automatically learn production-line knowledge from knowledge graphs of the relevant DTs, consolidate variables and interdependencies to improve accuracy, and deal with process (sequence) dependencies that existing ML models cannot handle.[2] Unlike most existing predictive ML models that achieve a low accuracy (that is, 45%–55%) analyzing manufacturing data, dcML achieves much higher predictive accuracy (that is, 80%–95%) in DM, which allow DM solutions to correct issues before they arise.

We are evaluating DM solutions powered by this novel research in food, steel, and composite manufacturing lines across Australia. For example, such a DT and dcML-based DM solution is currently improving the consistency of food products manufactured in Bega's Port Melbourne plant in Australia. This DM solution[3] includes DTs for evaporators, separators, and pumps in Bega's plant, and a dcML model that predicts the consistency of food products (for example, vegemite and peanut butter) before they are manufactured. This dcML model has been trained with one year of production data collected from machine, products, materials, and plant operators to "understand" the dependencies and physical constraints between variables, such as machine settings and parameters. Such predictions allow timely automatic adjustments improving the consistency of the product by 17% in the first pass, and 100% in the second pass.

We are evaluating DM solutions powered by this novel research in food, steel, and composite manufacturing lines across Australia.

We are currently developing similar DM solutions for reducing machine breakdowns in steel manufacturing at Infrabuild's Laverton plant, and for improving the production efficiency of composite airframe parts via reducing their cure cycle in Boeing Australia. These DM solutions are achieving greater than 80% accuracy. We envision that future DM solution powered by DTs and dcML will improve production efficiency, enhance product quality, and reduce breakdowns by 20%–30% across the entire manufacturing industry. C

Digital Twins and dependency/constraint-aware machine learning supporting production improvements in food, steel, and composite manufacturing.

References

1. Bamunuarachchi, D., Georgakopoulos, D., Banerjee, A. and Jayaraman, P.P. Digital twins supporting efficient digital industrial transformation. *Sensors 21*, 20 (Oct. 2021), 6829, MDPI.
2. Bamunuarachchi, D., Georgakopoulos, D., Jayaraman, P.P., A Banerjee, A. A framework for enabling cyber-twins-based industry 4.0 application development. In *Proceedings of the 2020 IEEE Intern. Conf. on Services Computing* (Sept. 2021).
3. Banerjee, A., Forkan, A.R.M., Georgakopoulos, D., Karabotic Milovac, J., and Jayaraman, P.P. An IIoT machine model for achieving consistency in product quality in manufacturing plants, (Sept. 2021), arXiv:2109.12964.
4. Piroumian, V. Making digital twins work. *IEEE Computer 56* (Jan. 2023), 42–51.

Dimitrios Georgakopoulos is a professor and Director of the Key IoT Lab, and Industry 4.0 Program leader at Swinburne University of Technology, Melbourne, Australia.

Prem Prakash Jayaraman is a professor and Director of Factory of Future and Digital Innovation Lab at Swinburne University of Technology, Melbourne, Australia.

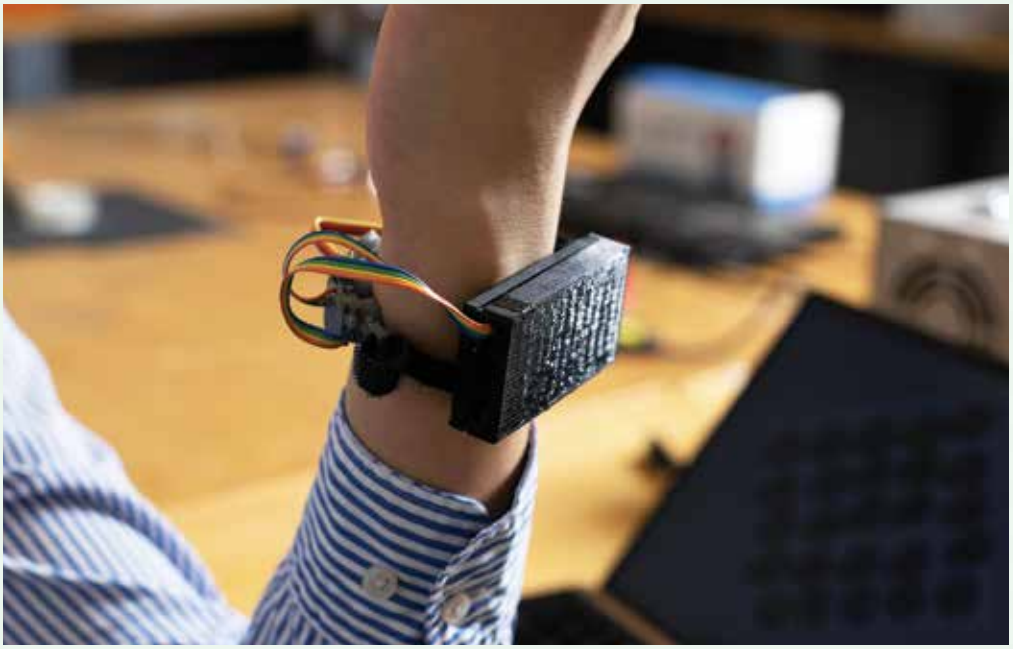
Co-Design Tools | DOI:10.1145/3589150

Co-Designing Personalized Assistive Devices Using Personal Fabrication

BY ANUSHA WITHANA

ASSISTIVE OR ENABLING technologies aim to create more accessible and inclusive solutions for people living with disabilities. This is critical, since many such users rely on technology for daily activities such as mobility and communication. While the problems are global, there are unique challenges that exist in the Asia Pacific region when it comes to developing assistive technologies, particularly assistive devices.

The United Nations Economic and Social Commission for Asia and the Pacific (UN ESCAP) estimates that 650 million people in the Asia-Pacific region live with a disability.[4] It is also well understood that disability statistics in the region could be significantly underreported. An Indonesian study published by TNP2K found that accessibility to equal opportunities, environments, education, and employment is significantly modest compared to many developed countries. Similarly, UNICEF "Every Mind" program in Sri Lanka reported that children with learning disabilities are often excluded from mainstream education. In some areas, sociocultural beliefs and taboos are associated with disability leading to social exclusion. This is overshadowed by the economical and geographical challenges.



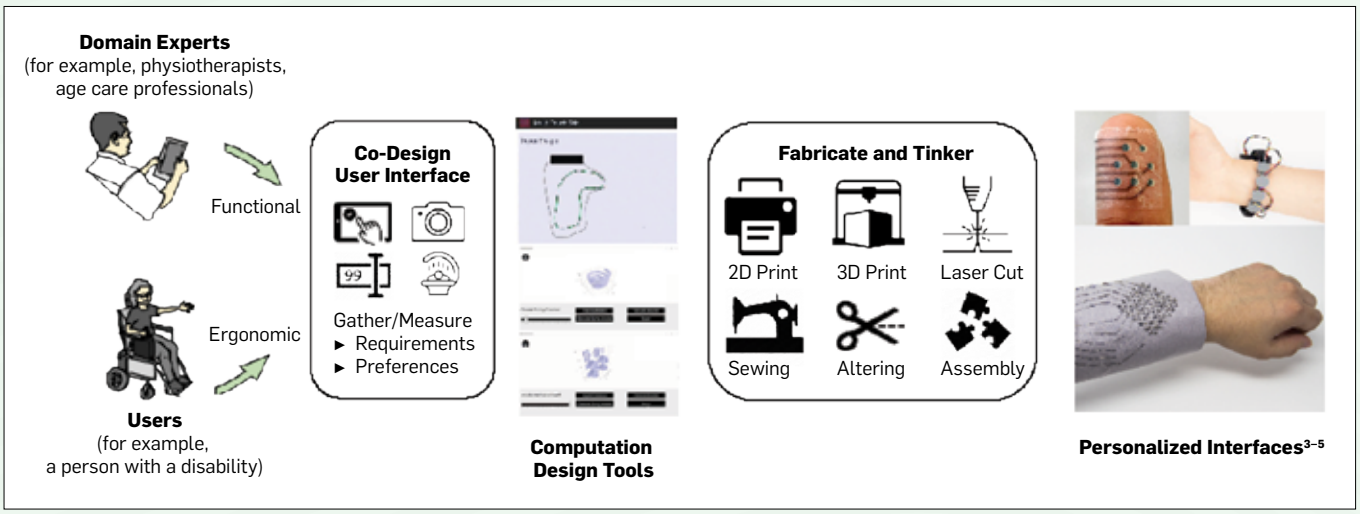
Researchers at the University of Sydney designed a 3D-printed sensor bracelet to give back control to the hand-impaired.

PHOTO COURTESY OF THE UNIVERSITY OF SYDNEY

The great diversity of people living in the region and the inherent differences in the way disabilities manifest in individuals create a demand for personalized solutions.

Inaccessibility to enabling environments and assistive technologies has been identified as a major contributor to the exclusion of people with disabilities in the region. This is further prevalent in rural and underpopulated areas where even in developed countries like Australia and New Zealand, access to technologies and specialists can be difficult to find. Some important challenges in the region include:

Need for personalized solutions. The great diversity of people living in the region and the inherent differences in the way disabilities manifest in individuals create a demand for personalized solutions. It is reported that a major reason for technology abandonment is the "poor fit between the user's abilities and the system's characteristics."[2] On top of the compulsory functional requirements, we learn that the social acceptance of technologies creates a big impact on the user experience. For instance, wearable technologies such as smart textiles and jewelry tend to mimic mainstream fashion accessories. In the Asia-Pacific region, fashion is so diverse and intertwined with cultures, there must be a mechanism for



Design Pipeline: Computational design tool for co-designing devices and printing them through additive manufacturing.

personalizing devices to different regions as well as to the individual.

Affordability. Mass-manufactured technologies leverage the economies of scale, or production at higher quantities to reduce the cost. Markets for enabling technologies are significantly smaller. Therefore, the unit cost of enabling technologies is much higher. Home to many developing countries, people living with disabilities in the Asia-Pacific region find the affordability of technology a major barrier to adaptation.

Access to experts. The region covers many remote, low-population-density areas. For instance, Mongolia and Australia are among the 10 least-populated countries in the world. Access to experts who can help with identifying solutions and making customization needed to fit the user is lacking in many areas, especially in terms of providing timely consultations.

In our research, we examined the cooperative design, or co-design, space of additive and subtractive manufacturing to create personalized assistive technologies.[1,3,5] For instance, 3D printing, laser cutting, and conductive inkjet printing can be combined with easy-to-use computational tools to create accessible approaches for non-tech-expert users to create personalized solutions. This is collectively referred to as *personal fabrication*—a unique solution to situations where mass production, that is, "designed for many," can be replaced with personal fabrication, that is "designed for the individual."

Our recent research shows that interactive computational tools can be used to personalize wearable devices, their functionality, and the form factor including the look and feel.[1,3] As shown in the accompanying figure, these tools can be utilized to capture the user needs and the expert recommendations to make the devices appealing to the unique needs of the individual[5] and significantly improve the quality of operation to fit the abilities of the user.[1] For example, if a smart textile solution is needed, it should be easily customizable to be woven into a saree in India and a batik in Indonesia. Not only will this approach be a game changer to fit the unique needs of the user, but the devices can also be easily modified by reprinting to fit changing requirements.

Furthermore, with the increasing affordability of additive manufacturing technologies such as 3D printers, we envision this technology will be available locally, taking the supply chain and logistic issues of rural areas and developing countries out of the equation. For instance, we now have remote sites in Sri Lanka where small-scale and affordable fabrication setups are replicated. We believe that personal fabrication can bring a scalable solution to the challenges of creating accessible technology in the region. Additive manufacturing also elevates the impact of economies of scale since they are always uniquely manufactured.

Finally, since the software tools can operate over the Internet, users and experts can co-design the devices online or remotely and print them locally. This increases access to experts in regional areas.

Our current research funded by the Australian Research Council and the Cerebral Palsy Alliance Australia looks at solving the technical challenges of creating computational co-design tools where an expert in Australia can assist in the development of a personalized device for a child living with cerebral palsy in Sri Lanka. A user with such experience welcomed the approach and commented it is "promising in terms of functionality and affordability." C

References

1. Lin, S.S. et al. MyoSpring: 3D printing mechanomyographic sensors for subtle finger gesture recognition. In *Proceedings of 2022 TEI Conf.*; https://doi.org/10.1145/3490149.3501321
2. Myrden, A. et al. Trends in communicative access solutions for children with cerebral palsy. *JCN 29*, 8 (2014), 1108–1118; http://dx.doi.org/10.1177/0883073814534320
3. Nittala, A.S. et al. Multi-Touch Skin: A thin and flexible multi-touch sensor for on-skin input. In *Proceedings of ACM CHI 2018 Conf.*; https://doi.org/10.1145/3173574.3173607
4. UN ESCAP. Disability in Asia and the Pacific: The Facts (2016); https://bit.ly/3Z2vgzU
5. Withana, A. et al. Tacttoo: A thin and feel-through tattoo for on-skin tactile output. In *Proceedings of 2018 UIST*; https://dl.acm.org/doi/10.1145/3242587.3242645

Anusha Withana is an assistant professor and ARC DECRA Fellow (DE200100479) in the School of Computer Science at The University of Sydney, Australia.

Wearables | DOI:10.1145/3592539

JIZAI Body: Human-Machine Integration Based on Asian Thought

BY MASAHIKO INAMI

"JIZAI BODY"[2] ALLOWS each person to live as they wish by controlling their natural body and altered robotic body parts. The Japanese term "*jizai*" originates from the Sanskrit word "*isvara*," which means supreme being free from earthly desires and constraints. JIZAI Body emphasizes the transformability of physical bodies and the effects on spiritual and internal minds fostered in Asian spiritualities such as Buddhism, Hinduism, and others, beyond enhancement or augmentation contextualized in Western cultures.

Figure 1. JIZAI ARMS.

For example, a wearable robotic limb system called JIZAI ARMS[4] (see Figure 1) allows social interaction between multiple wearers and explores communication between them. Wearers can exchange the arms and share the body parts. A robotic sixth finger (Figure 2) enables humans to embody a robotic finger independently from the innate fingers.[3] It has also been shown that humans can embody supernumerary robotic arms and feel them as their own limbs in a virtual reality environment.[4]

Figure 2. Sixth finger.

The uniqueness of JIZAI Body research lies in its investigation of what society will be like when people who acquire JIZAI bodies interact, and in its scientific elucidation of JIZAI states. It raises the question of how people's body image will change in the JIZAI society. Both the realization of JIZAI bodies based on Asian thought, and the scientific clarification of JIZAI states, is required. C

References

1. Arai, K. et al. Embodiment of supernumerary robotic limbs in virtual reality. *Scientific Reports 12*, 1 (2022), 9769; https://doi.org/10.1038/s41598-022-13981-w
2. Inami, M. et al. Cyborgs, human augmentation, cybernetics, and JIZAI body. In *Proceedings of the Augmented Humans Conf.* (Kashiwa, Chiba, Japan, 2022). ACM, 230–242; https://doi.org/10.1145/3519391.3519401 https://doi.org/10.1145/3544548.3581169
3. Umezawa, K., Suzuki, Y., Ganesh, G., and Miyawaki, Y. Bodily ownership of an independent supernumerary limb: An exploratory study. *Scientific Reports 12*, 1 (2022), 2339; https://doi.org/10.1038/s41598-022-06040-x
4. Yamamura, N. et al. Social digital cyborgs: The collaborative design process of the JIZAI ARMS. In *Proceedings of the 2023 CHI Conf. Human Factors in Computing Systems* (Hamburg, Germany, Apr. 23–28, 2023), ACM, New York, NY, USA;

Masahiko Inami is a professor in the Research Center for Advanced Science and Technology at the University of Tokyo, Japan.

Technical Inventiveness | DOI:10.1145/3589949

Fusing Creativity and Innovation in Asia's Manufacturing Industry

BY YOSHIHIRO KAWAHARA

JST ERATO "KAWAHARA Universal Information Network Project" is a unique research program that unites researchers in the fields of human-computer interaction (HCI), robotics, and electrical engineering in Japan.[1] The program aims to explore the technologies needed for the next generation of the Internet of Things and HCI.

The program's cross-disciplinary approach has brought unique research results through collaboration among different fields. Since the Asian region, including Japan, is a unique global manufacturing hub, this crossroads brought together experts from various fields of manufacturing, including top fashion designers, architects, and manufacturers of building machinery and materials.

One such notable success story is found in wireless power transfer technology. We devised a new structure called the Multi-mode Quasistatic Cavity Resonator to create a space that can transmit tens of watts of power anywhere in a three-meter square room.[2] The key to this new structure is the resonance conditions that allow the walls, floor, and ceiling of the room to operate as low-loss power transmission resonators that are computationally determined. As a result, it is possible to generate a three-dimensionally distributed AC magnetic field that fills the entire room. We also developed Meander Coil++, a garment that can feed several watts of power to a device on the body by designing a knit with low-loss coils arranged in a zigzag pattern to reduce the magnetic field in the body.[3] Our access to an automatic knitting machine is an important component of this invention.

Typically, research in this field has been focused on electrical engineering. However, the project explored the human experience of wireless power transfer and considered how to create a system that blends into the environment seamlessly and without discomfort, as a part of building materials or clothing. This user-centered approach represents a departure from traditional technical-focused research and has the potential to greatly enhance the usability and convenience of wireless power transfer technology.

We have discovered that having a shared vision for the future and a united approach to technology can lead to the development of unique technical innovations.

Developing soft robots required a unique control method and new materials.

Discovering the importance of softness. "Softness" was also identified as a key technical challenge for future robots during a project discussion. Developing soft robots required a unique control method and new materials. One example of this is the "pump-less" phase-change actuator, which uses a circuit printed on a sheet less than one millimeter thick to slowly contract and produce a force. The liquid that was key to the realization of this actuator was not for robotic use, but a material used for cleaning and cooling semiconductors. We have also developed a caterpillar-shaped soft robot that can move freely in places, such as on thin branches or wires, using a theory inspired from a real caterpillar.

Poimo, a soft inflatable mobility vehicle in harmony with people and the environment. The personal mobility vehicle "poimo" is the embodiment of the three themes of "energy," "actuation," and "fabrication." It is a next-generation vehicle for travel within a short distance. One of the key features of the poimo is that it can be easily deflated and made compact for transportation when not in use. The body of the poimo is made of a balloon-like chamber that is constructed using a drop-stitch structure. This allows the body to be soft and flexible, while still being strong and durable.

Poimo was designed to be a truly personalized vehicle, but because of its

Multi-mode Quasistatic Cavity Resonator uses a room's walls, floor, and ceiling as low-loss transmission resonators to send tens of watts of power anywhere in a three-meter square room.

light body, it is essential the body design take the rider's center of gravity into account. To this end, we developed a design tool that allows users to create a vehicle that fits their size simply by posing for a ride. The sofa type was well received by those who test drove the vehicle. Because of the spread of COVID-19, we received many inquiries from overseas as well. The idea's prototyping and customization would not have been possible without the cooperation of manufacturers who use the same material to make other products, including inflatable surfboards.

Conclusion

It can be difficult to establish shared technical objectives in a project involving a diverse group of stakeholders. However, we have discovered that having a shared vision for the future and a united approach to technology can lead to the development of unique technical innovations. The national project concluded in March 2022, but implementing it in the real world presents a new challenge for the project members. C

References

1. JST ERATO Kawahara Universal Information Network Project; https://www.jst.go.jp/erato/kawahara/
2. Sasatani, T., Sample, A.P., and Kawahara, Y. Room-scale magnetoquasistatic wireless power transfer using a cavity-based multimode resonator. *Nat Electron 4*, 689–697 (2021); https://doi.org/10.1038/s41928-021-00636-3
3. Takahashi, R., Yukita, W., Yokota, T., Someya, T., and Kawahara, Y. Meander Coil++: A body-scale wireless power transmission using safe-to-body and energy-efficient transmitter coil. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, Article 390, 1–12; https://doi.org/10.1145/3491102.3502119

Yoshihiro Kawahara is a professor at the University of Tokyo, Japan.

Health Care | DOI:10.1145/3589232

MEXT Society 5.0 Realization Research Support Project

BY TERUO HIGASHINO, HIDETOSHI KOTERA, AND YASUSHI YAGI

In recent years, innovations such as the Internet of Things (IoT), big data, robotics, and AI have become part of everyday life, helping people lead active, high-quality lives and creating a super-smart society. Society 5.0 was proposed by the Japanese government as a vision of a future society toward which Japan should aspire.[a] The society that Society 5.0 aims to develop is similar to the society the U.S. NSF's Smart & Connected Communities/Health projects aim to create. In 2018, Osaka University's "Initiative for Life Design Innovation (iLDi)" project was selected for the Ministry of Education, Culture, Sports, Science and Technology (MEXT) Society 5.0 Realization Research Support Program.[b]

The IoT has a variety of application domains, including health care.[1] The goal of our iLDi project is to create innovations to maintain people's health and to improve illness using ICT. We are building a Personal Life Record (PLR), a database that adds daily life activity data to a Personal Health Record (PHR), and by comparing this with everyone's medical information, we try to provide timely support appropriate to each person's health situation. While PHR is a database of human medical and health information, PLR incorporates data about various daily activities such as regular habits, diet, exercise, and workplace and school activities.

We are simultaneously building MYPLR—a platform for collecting PLR data—and creating mechanisms for secure management and secondary use of personal information. MYPLR complies with the EU's General Data Protection Regulation (GDPR) approach and has built-in capabilities to make health-related information obtained from individuals available for secondary use by organizations and companies seeking to employ that data. In the case of secondary use, MYPLR provides individuals with information on the organizations that wish to use their data and the purpose of use and has established a mechanism to provide information only upon their consent.

In our project, we are promoting "lifestyle" research aimed at maintaining and improving quality of life (QOL); "wellness" research aimed at improving mental and physical health; and "edutainment" research to appreciate enjoyment and learning, which aims to promote learning with high motivation for school attendance (as illustrated in the accompanying figure). Our goal is to design a life with high QOL based on physical, mental, social, (communication), and environmental health from the relationship between people and their everyday lifestyle, and to disseminate various technological innovations and changes in the socioeconomic environment from the university.

We are conducting the following four research projects for utilizing PLR data and validating its usefulness in real-world situations:

► **Health preservation/preventive medical care project.** The work of this group includes three projects: "Watch over 1,000 days of pregnancy and parenthood," "Elderly monitoring," and "Heart failure prevention." For example, in the first project, we are developing a system to classify the troubles and symptoms of pregnant women into several categories and provide timely advice and instructions according to each situation.

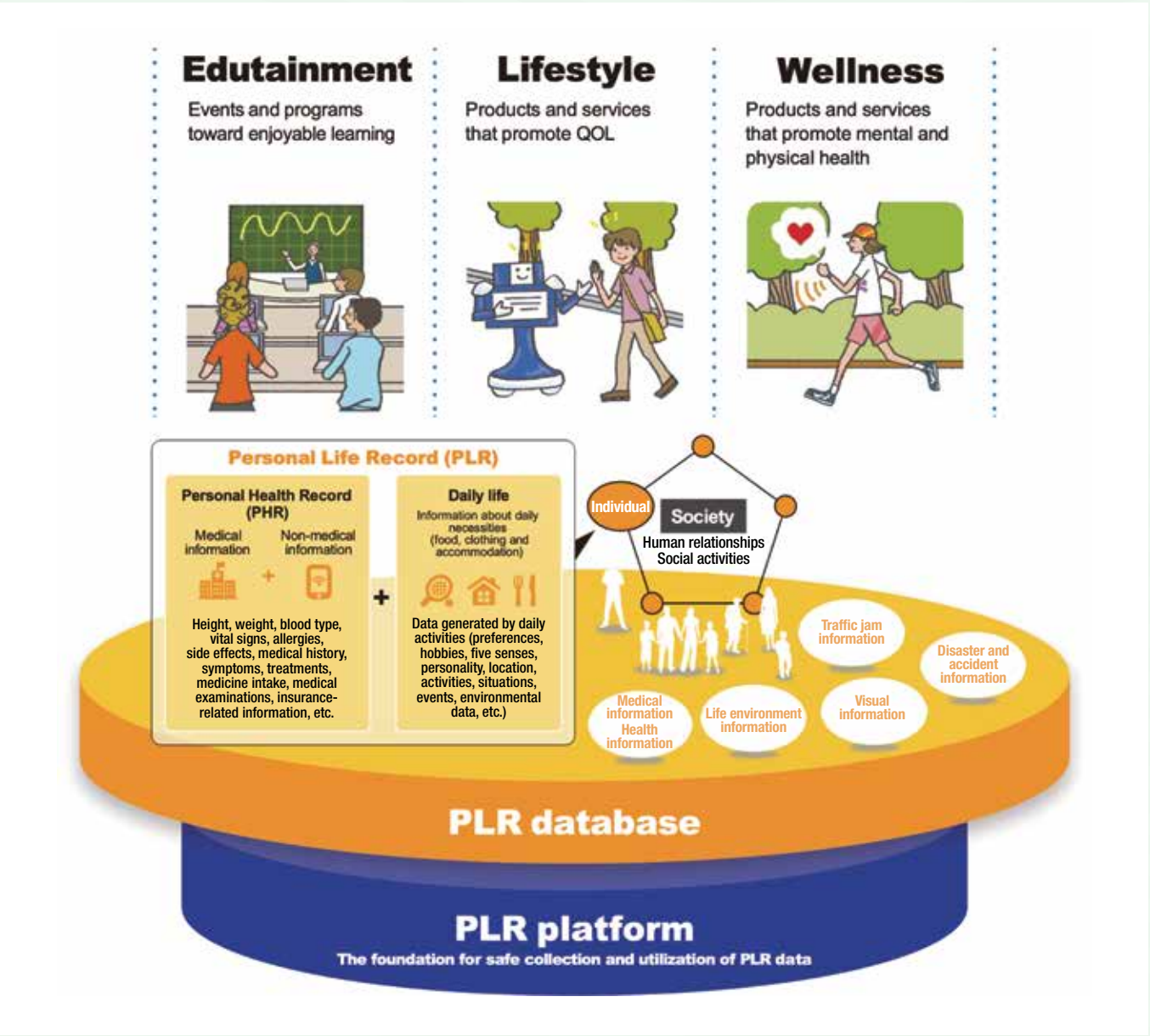
► **Health and sports project.** The work of this group includes two projects: "Preventing and predicting sports injuries" and "Detecting early signs of heatstroke." Based on muscle, ligament, and thermal models of the human body, we are building systems that use AI and simulation technology to estimate in real time the recovery status of the exercise function of the injured person and changes in the core body temperature of the exercise, and appropriately advise whether it is safe to continue exercise.

► **Future school support project.** The work of this group includes two projects: "Detection of early warning signs of Hikikomori" and "Education support and student life support." Those projects use smartphones and IoT devices to estimate students' learning and living conditions, advise them on how to improve their learning efforts and lifestyle habits, and recommend whether they should visit the University's health and counseling center.

► **Symbiotic intelligent system project.** The work of this group includes two

Society 5.0 was proposed by the Japanese government as a vision of a future society toward which Japan should aspire.

a https://www8.cao.go.jp/cstp/english/society5_0/index.html
b https://www.ids.osaka-u.ac.jp/ildi/en/index.html



Three goals of the MYPLR project.

projects: "Elderly healthcare stimulating conversations using an interactive robot" and "Creating a comfortable environment via interactive robots and environment control." Those projects use android robots to promote conversation and improve the living environment of elderly people.

In addition, four social implementation projects are underway to enable the collection and use of PLR data. The socio-technical project addresses issues related to privacy and ethical, legal, and social implications, and explores ways to enable the widespread use of PLR in the real world. We also focus on professional development and have invited research proposals from young researchers.

The basic structure of each technology was created in the first five-year research period starting in 2018, and we plan to work on the social implementation of these technologies in the second research period that started last April. So far, we have created various sensing mechanisms. During this period, we look to combine PLR data with personal medical information to create mechanisms to induce appropriate behavioral changes based on sensing results according to the individual's situation, thereby creating innovations useful for improving and promoting good health. Since healthcare data is massive and pertains to racial and cultural backgrounds, the plan is to build AI chatbots suitable for such behavioral changes. We also plan to contribute to the Osaka Expo to be held in 2025. C

Reference

1. Islam, S.M.R. Kwak MD., D., Kabir, H., Hossain, M. and Kwak, K-S. The Internet of Things for health care: A comprehensive survey. *IEEE Access 3* (2015), 678–708.

Teruo Higashino is Specially Appointed Professor at Osaka University, Japan and Professor/Vice President at Kyoto Tachibana University, Japan.

Hidetoshi Kotera is Specially Appointed Professor at Osaka University, Japan.

Yasushi Yagi is Professor at Osaka University, Japan.

DOI:10.1145/3597464

Article development led by acmqueue
queue.acm.org

The loss of potential gain from other alternatives when one alternative is chosen.

BY KELLY SHORTRIDGE AND JOSIAH DYKSTRA

Opportunity Cost and Missed Chances in Optimizing Cybersecurity

CYBERSECURITY INVOLVES EVERYDAY decisions about balancing costs that influence defensive outcomes—where to focus resources, which threats pose the most critical impact, and which mitigations must be deployed before others. Problems abound in this endeavor. Cost is not just monetary; resources are finite and scarce. Consequently, cybersecurity decisions are in danger of suboptimal outcomes and missed opportunities.

In theory, decisions should be made relative to the expected returns on each option. For example, will backups protect against the expected losses from ransomware? Other alternatives are, by necessity, not pursued. The calculation of ROI (return on investment) determines the value of a particular choice but ignores what might have been.

This is the very definition of opportunity cost: the loss of potential gain from other alternatives when one alternative is chosen.[6] The money spent on data backups cannot also be used for endpoint protection as a defense against ransomware.

Cybersecurity has much to gain by incorporating opportunity cost into decision making, from minimizing the impact of external threats to maximizing company productivity. If resources are spent defending against supply-chain attacks when social engineering results in larger or more frequent losses, the business suffers. Inside the company, allocating time and other costs to security can even be detrimental to the bottom line.

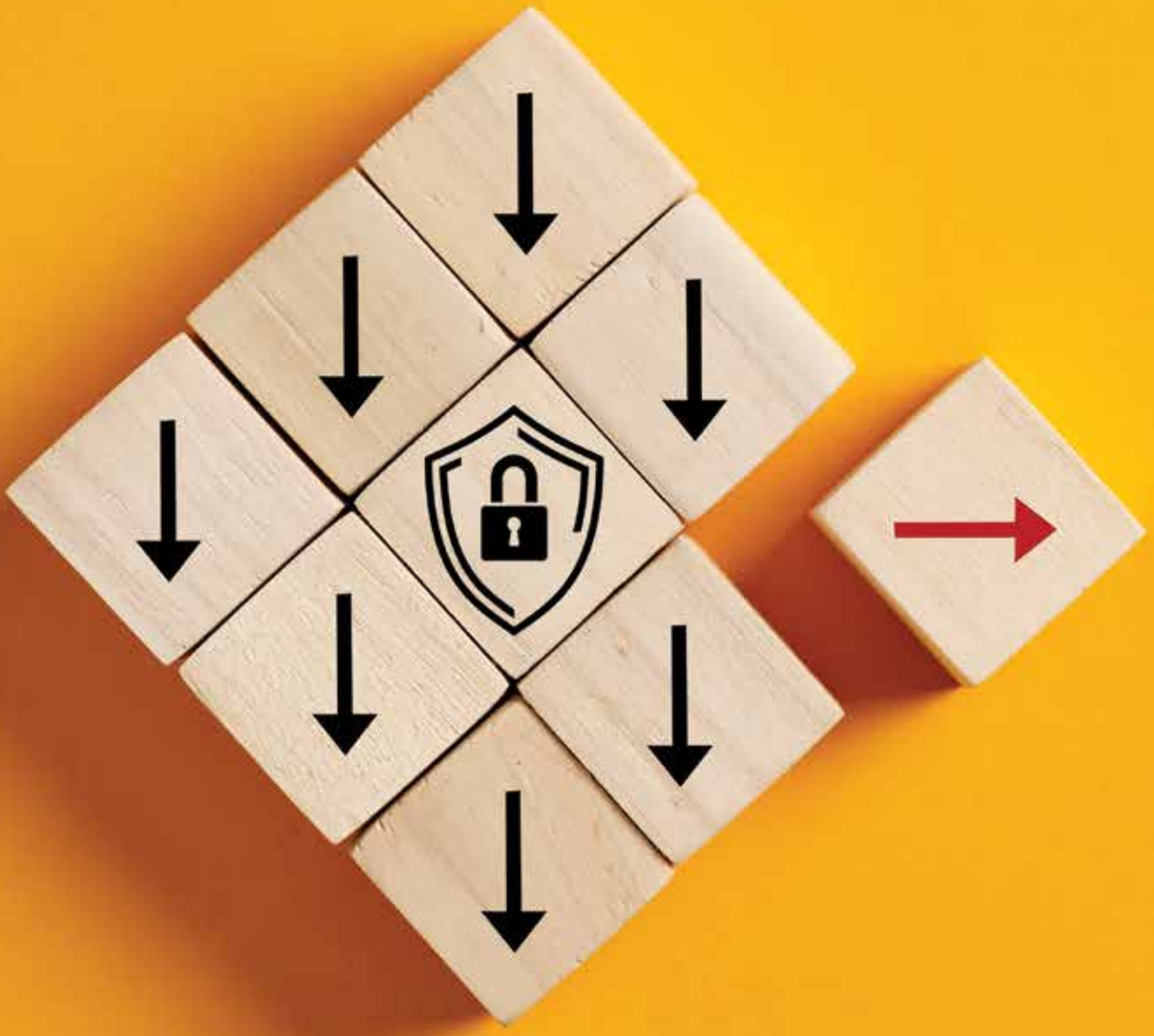
This article explores how opportunity cost is commonly neglected in practice and discusses ways to incorporate it into routine decision processes; a case study illustrates the benefits of making opportunity cost a core component of cybersecurity decision making.

A frustration for many people is calculating the expected benefit of a given choice. What is the value in spending four hours reading a new book rather than practicing the violin? Considering alternatives is not a default human tendency, but these are not impossible questions to answer, even absent absolute dollar values.

Cost is usually associated with monetary value. Financial resources spent on one opportunity means those resources are not available for something else. Businesses calculate the financial cost of various components, including licensing, capital expenses, and the fully loaded cost of labor.

The monetary cost and benefit of a specific choice are often considered in isolation, rather than taking into account its *externalities*—the cost

IMAGE BY CAGKAN SAYIN



and benefit imposed on other entities by the choice.[12] A classic example of a negative externality is pollution. A company can perform a cost-benefit analysis to calculate its optimal production rate, but this will consider only the cost incurred to make a widget and the benefit the company receives by selling the widget. It will not consider the costs imposed on the local community by lowering the quality of drinking water, nor the costs imposed on society through carbon emissions. In cybersecurity, implementation of an application security tool may present a positive ROI for the security team but may also result in a negative externality of slower or fewer software releases, imposing a cost on the software engineering team(s), the organization, and its customers.

Table 1. Costs of opportunity cost in cybersecurity.

Tangible Costs	
Organizations	▸ Wages of security personnel
Intangible Costs	
Employees	▸ Time investment in security intervention ▸ Burnout (lack of meaning or perceived lack of impact) ▸ Cognitive overload (task switching) ▸ Lack of perceived "flow" (wait time, number of interruptions) ▸ Cynicism, exhaustion ▸ Workplace conflicts / friction
Organizations	▸ Productivity loss from implementing intervention ▸ Time to plan and execute intervention ▸ Delayed time to market ▸ Increased attrition ▸ Curtailed innovation
Society	▸ Consumer workload and anxiety ▸ Efficiency loss of public funds ▸ False sense of security ▸ National security risk from third parties (defense industry) ▸ Research costs

Table 2. Effects of opportunity cost in cybersecurity.

Productivity Effects	
Employees	▸ Satisfaction (perception of work) ▸ Efficacy (availability of resources needed to complete their work) ▸ Less burnout and stress ▸ Collaboration, reduced conflicts ▸ Efficiency and perceived flow ▸ Standardized, less manual work (fewer mistakes, less toil) ▸ Faster. empowering onboarding
Organizations	▸ Product quality gains (reliability, service health, number of bugs) ▸ Faster time to market ▸ Incident reduction (impact / severity, volume. duration) ▸ Speed of change integration ▸ Turnover and attrition reduction ▸ Learning culture (innovation, knowledge discoverability) ▸ Activity volume (PRs, deploys, infrastructure utilization)
Society	▸ More dependable digital services ▸ Macroeconomic effects ▸ Reduced identity theft and fraud
Tangible Benefits	
Organizations	▸ Revenue growth ▸ Customer satisfaction (renewals, expansion, feature adoption) ▸ Output gains (more software releases, more product lines) ▸ Profitability (doing more with less) ▸ Uptime (service availability)
Intangible Benefits	
Users	▸ Enhanced user experience ▸ Reduced anxiety / worry about safety and privacy
Employees	▸ More dedication, less cynicism ▸ Feeling that their work is meaningful and has impact ▸ Energized vs. drained
Organizations	▸ Brand perception and corporate image ▸ Competitive advantage ▸ Compliance adherence ▸ Intellectual property ▸ Talent attraction and retention

The costs and benefits are likely to be quite different from the perspective of a security team, the organization and other teams within it, the users and customers, and the society around them. A benefit for one stakeholder may beget a cost for another. Tables 1 and 2 untangle this complexity by illustrating the costs and effects of opportunity cost in cybersecurity. It distinguishes tangible from intangible costs/benefits and highlights that these apply to several key groups: employees, organizations, and society.

Time is an important component of opportunity cost but is often neglected in practice. It takes time to answer help-desk tickets and to develop new security policies. The investment of time also has second-order impacts, on employee satisfaction, burnout, turnover, and more. Emotional experience, such as anxiety, frustration, and confusion, is often overlooked as a cost when evaluating options, despite its well-documented salience during human decision making.[10] The stress of time pressure can exacerbate users' frustrations with a security tool's lack of usability, making them more likely to bypass the security requirement.[8]

Security programs are components of organizations and can expend energy or absorb it, but energy is neither created nor destroyed. Beseeching employees to be vigilant to phishing threats requires them to expend energy, which the security team absorbs (as these user efforts allow the security team to expend energy elsewhere). Requiring software engineers to triage bugs discovered by vulnerability scanners is another example; developers expend energy combing through findings and fixing them, and the security team absorbs that energy. Thinking of where energy is expended and absorbed, and by whom, can help excavate the opportunity cost of a security decision.

Salience and the Null Baseline

The most effective method of encouraging exploration of opportunity cost

during decision making is to make alternatives more salient.

How can alternatives outside the focal point be made more salient? One way is to require a baseline to *do nothing*—that is, the costs and benefits of an option must be compared with those of doing nothing. This is known as the *null baseline*. The idea is to encourage salience around alternatives that solve the same goal. For example, if your goal is relaxation, you should compare not only vacation options, but also local spa days, outsourcing domestic toil such as cleaning, or implementing a yoga routine. This comparison should involve not just monetary and effort costs, but also benefits in terms of positive memories, health outcomes, increased leisure time, and whether the benefits are acute or extended.

Focal options influence security consumption. For example, security teams will often evaluate products of the same type, known as a *bake-off*, which reflects the focal options. They are less likely to include nonfocal alternatives during the same decision making process. Nor are they likely to compare the costs and benefits of buying the focal options vs. not buying any of them at all.

A security team may select three security-awareness training vendors for a bake-off to compare their costs and benefits. The salient costs might include the upfront purchase price and an estimate of initial implementation costs. The salient benefits might include the volume of content, variety of content, and customer support.

The costs and benefits that are *not* salient are those outside of the focal options and local context of the security team. The cost of internal end users disrupting their work to perform the training is not salient. The cognitive cost of users—or the security team itself—viewing the problem as "solved" when it remains unsolved is not salient.[9] The benefits of pursuing alternative options—such as using single sign-on and multifactor authentication to reduce the impact of users falling for phishing schemes—are not salient. Neither are the benefits of simply not purchasing the tool, which could include better goodwill for the security team.

The most effective method of encouraging exploration of opportunity cost during decision-making is to make alternatives more salient.

Always considering a null baseline is a worthy heuristic for encouraging security decision makers to consider opportunity cost. This heuristic simplifies the consideration of opportunity cost and makes best use of finite time and attention. When evaluating a solution to a problem area, security professionals should consider "do nothing" as their baseline and approximate its benefits.

Suppose a security leader wishes to evaluate application-security testing tools. By considering the null baseline, the leader might come up with benefits including faster lead time for changes, increased deployment frequency, lower change failure rate, and regained time that would otherwise be diverted to triage test findings.[3] The opportunity cost of pursuing application security testing may be considered too great of an impact on the business relative to the benefit of fewer bugs reaching production (which relies on a hypothetical outcome that those vulnerabilities will be exploited). If the security leader considers only the focal options, comparing the relative costs and benefits of each tool, these facets of the potential decision outcomes would be neglected despite bearing a nontrivial cost to the organization.

The null baseline can also support reframing the problem statement toward more globally optimal outcomes rather than local maximums. If the security leader's motivation behind application security testing tools is reflected in the problem statement of, say, "Fewer vulnerabilities must reach production," then considering the "do nothing" option may prompt the leader to rethink this problem in light of speed being vital to the organization. That is, the organization's priorities will become more salient, resulting in a problem statement more aligned with them such as, "Vulnerabilities exploited in production must not impact business operations." This changes the focal area substantially.

Alternative hypotheses not being salient also impacts security decision making. For example, once a security leader decides to adopt a zero-trust architecture, they are less likely to gather evidence about alternative hy-

potheses (that something other than zero trust might solve the problem best) and may become overconfident in the singular, original hypothesis.[11] Making opportunity cost salient, such as by considering the null baseline, increases the likelihood that consumers will compare options across categories and benefit types rather than solely comparing the narrow competitive set reflected by focal options.[15]

In cybersecurity, the null baseline would therefore allow security leaders to perceive competition across solution categories and reconsider the ultimate problem being solved. No longer is it the narrow framing of, "What insider threat tool is best?" Now the question is, "What solution will most effectively reduce the impact of unwanted activity by an insider, given resource constraints?" This facilitates comparison between insider threat tools and identity controls or other options.

Finally, the null baseline is helpful as a trivial stopping rule. The pursuit of additional information and evaluation of alternative choices could continue indefinitely if not for a decision to halt the process. For those looking at opportunity cost for the first time, the decision maker considers the null baseline and then stops seeking other options. Once this becomes routine, other stopping rules can be considered.

Areas of Application for Opportunity Cost in Cybersecurity

There are many ways to apply opportunity cost to cybersecurity, as illustrated by the examples just mentioned. This section explores four common applications in more depth.

Development and solution selection. Security practitioners struggle to select the best solutions to their problems under budget and time constraints. The potential solution to a given problem can be expanded when considering options such as build or buy; manual or automated processes; tightly scoped MVP (minimum viable product) or a more complete initial feature set. While this reflects a slightly more expansive consideration of solutions, it still reflects a zoomed-in perspective.

An opportunity-cost framing exposes situations when the problem definition is overly prescriptive, which stifles alternatives that might be superior solutions. For example, "What static application security tool is best?" begets a narrow focal point and prescribes only static application security testing (SAST) tools. "How can we minimize the number of security bugs developers introduce into code?" is less tool-specific and might introduce the focal option of security training for developers, but is still anchored to a relatively narrow focal point.

Risk quantification can be informative, but the most important outcome is how the results influence subsequent decisions and actions.

Instead, "How do we minimize the impact of security bugs in code running in production?" allows an expansive list of potential solutions to best prioritize allocation of finite resources. Brainstorming with other teams could result in a broader list including not only SAST tools and training, but also ephemeral or immutable infrastructure, standardizing libraries and patterns so developers are less likely to make mistakes, architecting the system with strong isolation properties, or running security chaos experiments to expose how the system behaves in adverse scenarios.[16]

Opportunity cost can assist with decisions on when to outsource work or perform it internally. Most organizations have a defined purpose to fulfill; their missions are not to secure their endpoints or networks. Through this lens, any work that does not directly support the organization's fundamental purpose bears the opportunity cost of taking time, budget, and cognitive effort away from work that more directly fulfills its purpose.

If a retail corporation wants a resilient online store, the expertise for databases, ad placement, and content delivery may not be among its internal skillset. Instead, it could outsource the problem to a platform service provider, advertising technology firm, and content-delivery network whose employees are skilled at such tasks. The opportunity cost of trying to create and operate a successful online store is time and cognitive effort that could be better spent dealing with retail business logic that would deliver more value to customers. This is a considerable cost to pay.

Requirements and patterns. A routine part of cybersecurity practice is

the use of requirements and patterns to define expectations around qualities and behaviors.

Defining requirements and creating reusable patterns reflects consideration of opportunity cost (which is perhaps why they are notoriously neglected in cybersecurity programs). Documented security requirements limit the variation of technology and attack surfaces, which supports repeatability and maintainability. Repeatability and maintainability reflect nonfunctional requirements that may not be in mind during cost-benefit analysis within a narrow focal point, despite their benefits to system resilience.[2]

For example, a common manual task for security teams is to answer engineering teams' ad hoc questions about how to build a product, feature, or system in a way that will be approved (or, at least, not be blocked) by the security program. This is often considered a high-priority activity, as ignoring it leaves engineering teams "stuck." Manual responses to each query by the engineering team, however, bear an opportunity cost, as the time spent answering each is time that could be used elsewhere to fulfill the security program's goals.

Defining explicit requirements and giving engineering teams the flexibility to build their projects while adhering to those requirements frees up time and effort for both sides: Engineering teams can self-serve and self-start without interrupting their work to discuss and negotiate with the security team, and security teams are no longer as inundated with requests and questions, so they can perform work with more enduring value. As a recent example, Greg Poirier applied this approach to CI/CD (continuous integration/continuous delivery) pipelines, eliminating the need for a centralized CI/CD system while maintaining the ability to attest software changes and determine software provenance.[14]

Spending time thinking through how to implement standards while minimizing friction and writing down the resulting security requirements—which become another set of nonfunctional requirements in a software project—provides a higher ROI than common security "toil" that provides only acute value. The opportunity cost of engineering teams asking for security requirements for each project includes tangible costs such as time and delayed realization of revenue, as well as intangible costs such as divided attention and cognitive overload. Writing documentation around, for example, "Here is how to administer a password policy" means that each engineering team no longer must ask for requirements, reducing these costs while providing the benefits of repeatability and maintainability. When faced with what type of password policy to implement, engineering teams should be able to access documentation and understand the requirements. Importantly, to minimize the opportunity cost of ad hoc discussions, everyone should reference the same single document to avoid one-off definition of requirements and negotiation based on individual tech stacks.

Therefore, the opportunity cost of not standardizing security requirements and disseminating them is high, but this cost is overlooked when practitioners consider the cost-benefit analysis of other work that takes time away from pursuing this standardization.

Risk prioritization. Despite much attention, there is insufficient analysis about the ROI and opportunity cost of risk quantification. The time cost of implementing information collection, tuning quantitative models, and interpreting information is often omitted; these activities explicitly siphon time from performing the activities that would address the concerns they attempt to measure. Calculating the probability of a particular security event sacrifices time that could be spent ensuring the organization is prepared for that security event and can minimize its impact. Risk quantification can be informative, but the most important outcome is how the results influence subsequent decisions and actions.

There is also the opportunity cost of delaying decision making because of the desire to calculate the probabilities of security events. Possessing a rough sense of what is easiest for the attacker—that is, the actions they are likely to attempt first in their operations—suffices to inform prioritization of effort. The benefits (in terms of real security outcomes) of refining those estimates with complex statistical models remain unproven.

Risk quantification can also lead to a false sense of security. By reducing uncertainty, practitioners may feel that the situation is more under control, but the ambiguity of the situation remains and can only be uncovered during real operations when real events occur.[17] This is true even beyond the realm of cybersecurity. As seismologist Susan Elizabeth Hough noted, "A building doesn't care if an earthquake or shaking was predicted or not; it will withstand the shaking, or it won't."[5] That is, the opportunity cost of attempting to predict an event is the time that could be spent cultivating resilience to the event.

Opportunity cost of delay. In cybersecurity, decisions related to waiting or proceeding are common:[1] Should we release or delay? Should we go for the temporary fix or wait for a more perfect solution?

Developing internal projects that aim for perfect and all-encompassing rather than a "good enough" MVP bears a similar opportunity cost. This is often witnessed as the dark side of the desire for automation. An organization may wish for "one automation framework to rule them all," even if it has not delivered anything after years of dedicated development work (exacerbated by the sunk cost fallacy). The opportunity cost for building the "perfect" thing for all use cases one year from now is solving one use case "good enough" sooner and delivering value sooner.

Security and software engineering teams alike often err in attempting to create automation frameworks that can be generalized to all potential use cases, often taking years to complete. This approach neglects opportunity cost, which would encourage comparison with alternative options, such as spending the same time, attention, and money on building automation for one specific use case to launch the project. For example, a security team could interview internal stakeholders and determine that automation for asset inventory would provide immediate value. They could constrain their

time and effort by meeting a set of minimum requirements for asset inventory automation, quickly building an MVP to solicit initial feedback and inform what work should come next.

The security engineering team can consider the opportunity cost of this subsequent work, too; perhaps the marginal benefit provided by requested features is less than other tasks the team could perform. The team can thereby reallocate its time to other tasks once the MVP is released rather than confining it in perpetuity.

Multiple smaller releases targeted at a few use cases fosters more flexibility when allocating time and effort, especially when it can take time for users and stakeholders to provide feedback on releases with enough clarity and consensus to inform next steps. Attempting a single massive release satisfying the ideal requirements not only makes time and effort an illiquid resource, but can also require other team members to work overtime to cover for the people building the Rube Goldberg machine. This leads to tangible outcomes such as poorer work quality, worse metrics, and increased attrition, as well as intangible outcomes such as declines in productivity, increased burnout, and climbing resentment. (The SPACE Framework for developer productivity suggests that software quality can be captured by measuring reliability, ongoing service health, and absence of bugs. These are all quantifiable metrics and the former two, being relevant for uptime, can be tied to revenue.[4])

Application and Practice in Cybersecurity

Here, we explore cybersecurity opportunity cost in practice, beginning with a detailed case study of Twilio, which highlights the principles discussed thus far. We then present tools to arm practitioners and security decision makers in implementing opportunity-cost consideration with maximum ease in real-world situations.

Case Study: Twilio is a technology company offering APIs and services for communications, including phone calls and text messages. For example, Uber uses Twilio to send text messages to customers about ride status. To make this possible, Twilio uses telecommunications protocols required for mobile network operations. To make this reliable, these protocols must be secured.

NightOwl is an automated testing framework built by two security architects at Twilio that performs attack-tree modeling of telecommunications protocols. Twilio gained a provisional patent for NightOwl, enhancing its intellectual property portfolio.

NightOwl saved $200,000 in its first year of use by just one team. Now multiple internal teams can use it to self-serve security testing ahead of product releases and are actively expanding their usage of it, proposing new features and functionality.

NightOwl's creators created automation and testing for two specific telecommunications protocols: GTP (GPRS Tunneling Protocol) and Diameter. These protocols are essential for operating Twilio's core mobile network, connecting users to Twilio's SIM cards across its global telecommunications network. Neither GTP nor Diameter support authentication or authorization, so testing each protocol's security and reliability is of paramount importance.

Contemplating opportunity cost led NightOwl's creators to scope an MVP that constrained its design to just these two protocols, ensuring the best use of their time and highest ROI for the organization. GTP and Diameter not only represented the first use cases for Twilio's Super SIM team, thereby offering the highest marginal benefits, but also were the easiest to build; one required a specialized transport layer, SCTP (Stream Control Transmission Protocol), and the other was over UDP (User Datagram Protocol).

Twilio's status quo was hiring an outside consulting firm to perform penetration tests for each protocol before a product became generally available (GA). Each test cost approximately $100,000 and required weeks to months of the security team's time not only to organize the tests with the consultants, but also to work through the findings afterward and coordinate with internal teams to fix them. Given the complex and specialized nature of telecommunications protocols, there were few off-the-shelf automated penetration testing tools that would present a direct alternative.

Because NightOwl's creators considered opportunity cost when pondering this problem, they understood that internal penetration testing was not viable. Hence, they expanded the range of viable alternatives to include developing an automated tool, trading off an initial upfront time investment for a reduction in ongoing costs and increase in perpetual benefits. By building their own tool, they could implement the comprehensive testing needed for such a complex system. There was more benefit in crafting targeted messages based on Twilio's real-world implementation than attempting to retrofit existing tools that covered only a subset of cases that were not applicable and did not cover other necessary tests. It also allowed them to implement new approaches that existing tools did not offer but that provided palpable marginal benefits, such as building a fuzzing engine and stateful messages to perform more in-depth tests.

Taking an MVP approach, which made best use of constrained time resources, allowed NightOwl's creators to deliver these benefits quickly while minimizing time costs. They arrived at the MVP-plus-iteration model by also considering the opportunity costs of delaying NightOwl's release, including not receiving user feedback, delayed delivery of value to users, and time that could be better spent on other projects.

The resulting savings from NightOwl for launching products to GA are both temporal and monetary. The first engineering team to use NightOwl was releasing a product to GA by a certain target date. The team ran the necessary tests for GTP and Diameter using NightOwl, leaving only one test for the outside consultants to cover. By covering testing for two out of three protocols, NightOwl saved $200,000. Had the other two tests not been satisfied by NightOwl, launching the product would have required an additional eight weeks of work. The engineering team even generated a report via NightOwl without assistance from the security team, including graphs of system outcomes when performing the simulated attacks. This pro-

vided not only the intangible benefit of gaining confidence in releasing the product to GA, but also knowledge documentation.

Before NightOwl, engineering teams would perform penetration tests only for major releases, given the substantial time, effort, and monetary expense. With NightOwl, engineering teams can now perform penetration tests before *every* release, including minor ones, thereby increasing the level of security coverage. The toil of coordinating the tests, creating tickets for each finding, and tracking completion of those tickets also imposed intangible costs—including stress, frustration, and lost productivity—eliminated by adopting NightOwl.

While it requires more effort upfront, NightOwl offers ongoing benefits that would have been missed had its creators not considered opportunity cost. They might have simply performed a cost/benefit analysis based on comparing different consultants or performing the same manual work internally, overlooking this superior option.

Finally, a consideration of opportunity cost inspired NightOwl's creators to reframe the problem statement. By considering the costs of interpreting the findings from penetration testing consultants, translating them for Twilio's engineering teams, and collaborating to prioritize fixes, they began thinking about the goal as "actionable security findings that engineers can fix on their own" instead of "conduct a penetration test." This refined NightOwl's initial functional and nonfunctional requirements, enabling its creators to deliver more value than the external penetration test or replacing it with manual scripting.

Implementing opportunity-cost consideration in practice. Any proposal to make decision making more burdensome risks skepticism and low adoption. Generating a list of alternatives might require research about which alternatives exist, the costs associated with different proposals, or even baseline measurements about the sum costs and value of a single option. Indeed, there is opportunity cost when considering opportunity cost.

Daniel Kahneman and other psychologists have differentiated fast and automatic thinking (system 1) from slower, analytical thinking (system 2).[7] Fast and automatic thinking—sometimes described as the "lizard brain"—is optimized for ease. The basic instincts of the human brain in the limbic cortex support survival and efficiency. Anything invoking slower, analytical thinking must feel to the thinker like it offers an outsized payoff, given the effort involved. Therefore, ease of implementation into decision making workflows is essential for opportunity-cost consideration to become the default. In real life, especially in fast-paced and stressful environments, people routinely have neither the time nor desire to think about more difficult decisions.

Opportunity cost is an integral aspect of cybersecurity that must be considered in every decision—a prominent, first-order component of decision-making, not an afterthought.

To incorporate opportunity cost in practice, decision makers should expend only enough time and brainpower to stimulate consideration of options beyond the focal point. The null baseline can serve as a new decision making heuristic—a mental shortcut that can lead to better decision outcomes—that can become more automatic with repetition.

In any given decision, you can try the following workflow to consider opportunity costs:

1. Consider the null baseline. What are the tangible and intangible benefits of doing nothing? What are the costs?
2. The null baseline's benefits become the starting point for considering the opportunity cost of the focal option.
3. The null baseline's costs may lead the decision maker to reframe the problem statement. For example, consider the focal option of manual security change approvals. The cost of "doing nothing" relative to this focal option is that a Web application running in production may be compromised. Thus, the problem statement should likely focus on "minimizing impact of compromise of code in production," for which manual security change approvals is one of many possible options.
4. With the reframed problem statement, consider the expanded set of options that could help solve it. View time, effort, and budget as fixed units interchangeable among options. Estimate the benefits and costs of each—

rough "napkin math" estimates are not only fine, but ideal here. Which option least erodes the benefits of the "do nothing" option? Answering that question can help uncover which option bears the most minimal opportunity costs.

These steps help the decision maker consider the null baseline, refine the goal, and explore the diversity of solutions.

There is no need to delve into precise opportunity-cost quantification; doing so could lead practitioners to become profligate consumers of time and energy. In essence, there is a nontrivial opportunity cost when quantifying opportunity cost with precision (if precision is even possible in cases where intangible costs abound). Each unit of time expended attempting to assign a specific number to opportunity cost in a decision area is a unit of time not spent implementing the right option based on existing analysis.

The greatest marginal benefits can be harnessed from considering opportunity cost by baselining against the "do nothing" option. The null baseline is likely to, at a minimum, make alternative goals salient—such as "delivering code to production faster," "being able to make more sales calls," or "being able to consume more data for analysis"—which can highlight what will be lost if a focal choice is pursued. The null baseline can also highlight otherwise overlooked intangible costs of the focal option, such as the impairment to productivity from engineers relearning access workflows when a zero-trust network access tool is adopted by the security team.

The null baseline is likely to encourage a reframing of the problem statement to expand the range of alternatives being considered as well—for example, reframing the statement "Buy the best code-scanning tool," to the more business-aligned and less myopic "Minimize the impact of bugs in production."

Once practitioners are comfortable with this opportunity-cost consideration workflow (that is, it becomes more automatic for the lizard brain), there are a few relatively thrifty options for measurement to complement this consideration. Pilot programs can uncover indirect and intangible costs toward considering opportunity cost before implementation. Proof of concepts of vendor solutions are often designed to hide implementation challenges, especially if teams affected by the tool are not involved in the bake-off. Piloting a solution, whether software or process, on one engineering team (with at least one other as a control) and comparing their goal metrics can generate evidence about the implementation's impact beyond the focal benefits and costs.

Another potential avenue for measuring opportunity cost is determining stakeholders' willingness to pay to *avoid* a security implementation. For example, how much would employees pay to avoid using the corporate virtual private network (VPN)? How much would developers pay to avoid using a SAST tool that takes six hours to scan their code or to avoid having to file a ticket to gain access to a service? Surveys about willingness to pay can serve to translate intangible burdens into tangible values, making it easier to compare benefits and costs across options.

Conclusion

Opportunity cost is an integral aspect of cybersecurity that must be considered in every decision—a prominent, first-order component of decision making, not an afterthought. Everyone in the cybersecurity ecosystem plays a role in security outcomes, no matter who is making decisions, from individual software developers to security managers to the CEO. The results will be the best possible options for security and privacy—ones that balance organizations' multifaceted goals to nurture resilient business operations in a complex digital landscape.

Security decision makers would benefit from recognizing the many types of costs in addition to money. They need not ignore opportunity cost for lack of precise measurements. In particular, one way to ease into considering complex alternatives is to consider the null baseline of doing nothing instead of the choice at hand. This can elucidate the "true" problem at hand.

Opportunity cost can feel abstract, elusive, and imprecise, but it can be understood by everyone, given the right introduction and framing. Using the approach presented here—especially the null baseline heuristic—will make it natural and accessible. The widespread inclusion of opportunity cost as a routine consideration in research and practice is a worthy goal. C

References

1. Arora, A., Caulkins, J.P., and Telang, R. Research note—sell first, fix later: Impact of patching on software quality. *Mgmt. Sci. 52*, 3 (2006), 465–471; https://pubsonline.informs.org/doi/10.1287/mnsc.1050.0440.
2. Cybersecurity and Infrastructure Security Agency, U.S. Digital Service, and Federal Risk and Authorization Management Program. CISA cloud security technical reference architecture, 2021; https://bit.ly/3IXmNIt.
3. Forsgren, N., Humble, J., and Kim, G. *Accelerate—The Science of Lean Software and DevOps: Building and scaling high-performing technology organizations.* IT Revolution Press, 2018.
4. Forsgren, N., et al. The SPACE of developer productivity: There's more to it than you think. *acmqueue 19*, 1 (2021), 20–48; https://dl.acm.org/doi/10.1145/3454122.3454124.
5. Hough, S. *Predicting the Unpredictable: The Tumultuous Science of Earthquake Prediction.* Princeton University Press, Princeton, NJ, 2010.
6. Huynh, T.N., Kleerup, E.C., Raj, P.P., and Wenger, N.S. The opportunity cost of futile treatment in the intensive care unit. *Critical Care Medicine 42*, 9 (2014), 1977–1982; http://bit.ly/3kObdat.
7. Kahneman, D. *Thinking, Fast and Slow.* Macmillan, 2011.
8. Kurowski, S., Fähnrich, N., and Roßnagel, H. On the possible impact of security technology design on policy adherent user behavior—Results from a controlled empirical experiment. *SICHERHEIT.* H. Langweg, M. Meier, B.C. Witt, and D. Reinhardt, eds. Gesellschaft für Informatik e.V., Bonn, Germany, 2018, 145–158; https://dl.gi.de/handle/20.500.12116/16276.
9. Lain, D., Kostiainen, K., and Capkun, S. Phishing in organizations: Findings from a large-scale and long-term study, 2021; https://arxiv.org/abs/2112.07498.
10. Loewenstein, G. and Lerner, J.S. The role of affect in decision making. *Handbook of Affective Science.* R. Davidson, H. Goldsmith, and K. Scherer, eds. Oxford University Press, Oxford, U.K., 619–664; https://bit.ly/3yh6X6s.
11. McKenzie, C.R.M. Taking into account the strength of an alternative hypothesis. *J. Experimental Psychology: Learning, Memory, and Cognition 24*, 3 (1998), 771-792; https://bit.ly/3ZAUENY.
12. Organization for Economic Cooperation and Development. Externalities—OECD. Glossary of statistical terms, 2003; https://www.oecd.org/regreform/sectors/2376087.pdf.
13. Podkul, C. Despite decades of hacking attacks, companies leave vast amounts of sensitive data unprotected. *ProPublica* (Jan. 25, 2022); http://bit.ly/3JhmFot.
14. Poirier, G. *Die softwareherkunft (software provenance): an opera in two acts.* Why would anyone do that? (Jan. 14, 2022); https://grepory.substack.com/p/der-softwareherkunft-software-provenance.
15. Russell, G. et al. Multiple-category decision making: review and synthesis. *Marketing Letters 10*, 3 (1999), 319–332; https://link.springer.com/article/10.1023/A:1008143526174#article-info.
16. Shortridge, K. and Rinehart, A. *Security Chaos Engineering: Sustaining Resilience in Software and Systems.* O'Reilly Media, Sebastopol, CA, 2022.
17. van Stralen, D. and Mercer, T.A. Ambiguity in the operator's sense. *J. Contingencies and Crisis Mgmt. 23*, 2 (2015), 54–58; https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-5973.12084.

Kelly Shortridge is a senior principal product technologist at Fastly, New York, NY, USA and co-author of *Security Chaos Engineering* (O'Reilly Media).

Josiah Dykstra is a technical fellow at the National Security Agency and the owner of Designer Security, LLC, Severn, MD, USA.

DOI:10.1145/3582491

While the success of early data-science applications is evident, the full impact of data science has yet to be realized.

BY M. TAMER ÖZSU

Data Science—A Systematic Treatment

THERE IS A data-driven revolution under way in science and society, disrupting every form of enterprise. We are collecting and storing data more rapidly than ever before. The value of data as a central asset in an organization is now well established and generally accepted. *The Economist* called data "the world's most valuable resource."[40] The World Economic Forum's briefing paper, *A New Paradigm for Business of Data*, states "At the heart of digital economy and society is the explosion of insight, intelligence and information—data."[5]

The field of data science is expected to enable data to be leveraged for making better decisions and achieving more meaningful outcomes. Although the term data science has some history, in its current incarnation as a modern field of study, it has already had significant economic impact. A 2015 Organisation for Economic Co-operation and Development (OECD) report identified "data-driven innovation" (DDI) as having a central driving role in 21st century economies, defining DDI as "the use of data and analytics to improve and foster new products, processes, organisational methods and markets." Data science deployments are still what might be called first generation, but their impact is already being felt in many areas: global sustainability,[11] power and energy systems,[25] biological and biomedical systems,[38] health sciences and health informatics,[12] finance and insurance,[8] smart cities,[33] digital humanities,[28] and more.

The last decade has established the terms "big data," "data analytics," and "data science" into our lexicon, both as buzzwords and as important fields of study. Interest in the topic, as evidenced by Google Trends (see Figure 1), has exploded over the same period. An increasing number of countries have released policy statements related to data science. In academia, data-science programs and research institutes have been established with significant speed, while many industrial organizations have created data-science units. A quick survey of these programs and initiatives suggests a common core but also a lack of unified and clear framing of data science.

There are several reasons why clarification is helpful. One is to be

» key insights

- **Establishing data science as a field of study or an academic discipline requires identification of its foundations and scope.**
- **Data science is not a proper subset of AI or of statistics. It is broader than data analytics using statistical or machine-learning techniques.**
- **Data science is highly interdisciplinary and involves STEM and non-STEM disciplines.**
- **Applications are crucial for data science, without which there is only big-data management and processing.**

IMAGE BY ARTHEAD

able to understand whether data science is an academic discipline. This is hard to know without a definition of data science and an identification of its core and scope. A related reason is to provide an intellectually consistent framing to the numerous data-science institutes and academic units being formed. A third reason is to bring some clarity to the question of who a data scientist is. The point is not to constrain what is meant by a data scientist or to limit the scope of current academic initiatives but to acknowledge the diversity around some commonalities. The final reason is that a systematic investigation of the field is likely to identify important techniques and tools that should be in a professional data scientist's toolbox.

Part of the difficulty is the carelessly interchangeable use of the terms "big data," "data analytics," and "data science" in much of the popular literature, which frequently spills over to technical literature. It is important to get them right. Data analytics, as defined in the next two sections, is a component of data science and not synonymous with it. Data science is not the same as big data. Perhaps the best analogy between them is that big data is like raw material; it has considerable promise and potential if one knows what to do with it. Data science gives it purpose, specifying how to process it to extract its full potential and to what end. It does this typically in an application-driven manner, allowing applications to frame the study objective and question. Applications are central to data science; if there are no applications to drive the inquiry, it is hard to argue that there is a data-science deployment. Jagadish also emphasizes this point, stating "'Big Data' begins with the data characteristics (and works up from there), whereas 'Data Science' begins with data use (and works down from there)."[22]

A second difficulty is the vagueness in many definitions about the relationship between data science, machine learning (ML), and data mining (DM)— this arises from the colloquial use of "data science" to mean data analytics using ML/DM. Data science is not a subfield of ML/DM nor is it synonymous with these disciplines. More broadly, data science is not a subtopic of AI—a common claim originating from confusion on boundaries. AI and data science are conceptually different fields that overlap when ML/DM techniques are used in data analytics but otherwise have their own broader concerns. The broader scope of data science is discussed in this article, highlighting its constituents that are not part of AI. Conversely, there are topics in AI, such as agents, robotics, automated programming, and others, that are not within the scope of data science. Thus, AI and data science are related, but one does not encompass the other.

A final difficulty is that data science is broad in scope, involving numerous disciplines, and finding the right synthesis is not straightforward. At the risk of oversimplification, the following are the different constituencies that have an interest in data science: STEM people who focus on foundational techniques and underlying principles (computer scientists, mathematicians, statisticians); STEM people who focus on science and engineering data-science applications (for example, biologists, ecologists, earth and environmental scientists, health scientists, engineers); and non-STEM people who focus on social, political, and societal aspects. It is important to include all these constituencies in discussions surrounding data science while establishing a recognizable core of the field. This is a difficult balance to maintain.

The objective of this article is to put forth an internally consistent and coherent view of data science. The article discusses core components, contributing disciplines, life cycle considerations, and shareholder communities. The main takeaways can be summarized as follows:

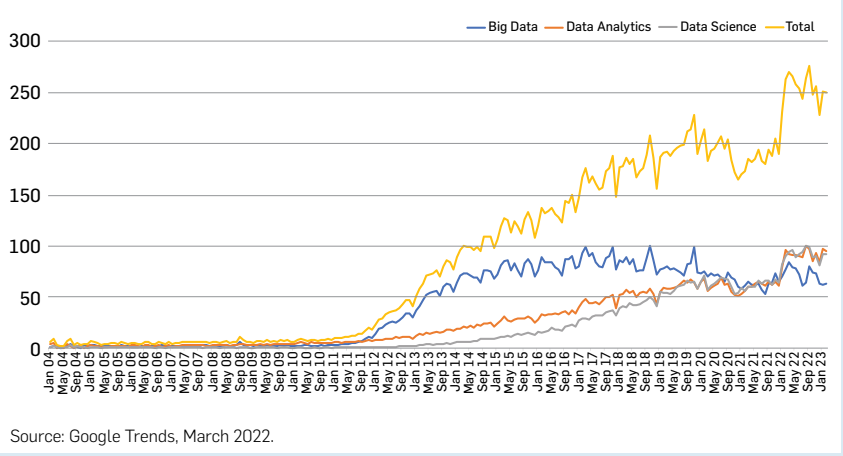
- It is important to clearly establish a consistent and inclusive view of the entire field, and one is proposed.
- To avoid becoming a catch-all or whatever particular circumstances allow, it is essential to define the core of the field while being inclusive, and four core areas are identified.
- It is critical to take a holistic view of activities that comprise data science.
- A framework must be established to facilitate cooperation and collaboration among a number of disciplines.

Data science is still in its early days as an emerging field. This article contributes to discussions around its nature and scope. There will, hopefully, be joinders to the discussion to better define the field.

What Is Data Science?

The origins of the term data science are fuzzy. Data is central to both statistics and computing, so both communities have tried to define the field. Statisticians suggest that its origins lead to John Tukey,[41] who passionately argued in the 1960s for the separation of "data analysis" from "classical statistics." His main point was that data analysis is an empirical science while

Figure 1. Trending of data science-relevant terms.



Source: Google Trends, March 2022.

classical statistics is pure mathematics. Tukey defines data analysis as "procedures for analyzing data, techniques for interpreting the results of such procedures, ways to plan the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." Capturing a precise definition of data has been important from the start of computing as a discipline. The International Federation of Information Processing's (IFIP) definition of data is "a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."[20] Naur builds on this definition: "Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences."[30]

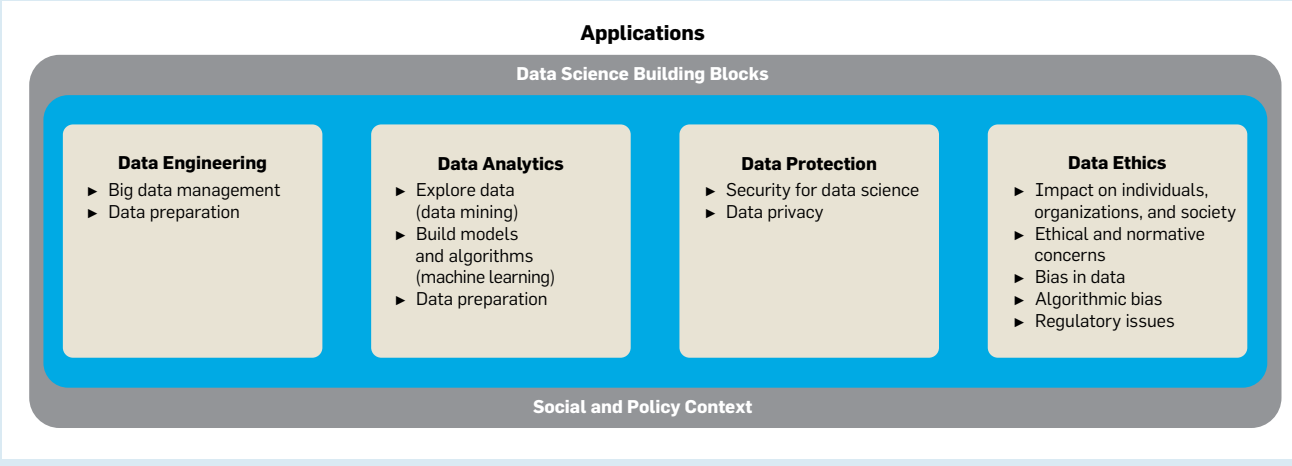
Clearly, both statisticians and computer scientists have been thinking about data science for a long time, and the understanding of what it is has evolved over time. A subtle shift happened in the 2000s with the recognition that data science is broader than data analytics and that it involves a process from data ingestion to the production of insights and actionable recommendations. During this period, data-intensive approaches to problem solving started to produce results. This became known as the 'big-data revolution' and resulted in data-centric approaches in many fields. What initially began as a computational paradigm (also called the third paradigm), where computational methods could replace or enhance laboratory experimental methods (a 2001 *New York Times* article declared "all science is computer science,"[23]) rather quickly changed to data-intensive methods. This is frequently referred as the fourth paradigm,[19] and data science systematizes this understanding.

There are significant differences between what was called data analysis (or analytics) and what the current understanding of data science entails. More modern definitions of data science encompass this broader interpretation—for example, "Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting non-obvious and useful patterns from large data sets."[26] The National Consortium for Data Science (NCDS), a U.S. consortium of leaders from academia, industry, and government, defines data science as the "systematic study of organization and use of digital data for research discoveries, decision-making, and data-driven economy."[2] These are not sufficiently precise to define a field, but they capture some important common themes: interdisciplinarity (as defined in Alvargonzález[4] and Choi and Pak[9]), a data-based approach to problem solving, the use of large and multi-modal data, the focus on deriving insights and value by discovering patterns and relationships in data, and the underlying process-oriented life cycle.

A working comprehensive definition that captures the essence of the field and explicitly recognizes that it involves a process would be: Data science is a data-based approach to problem solving by analyzing and exploring large volumes of possibly multi-modal data, extracting knowledge and insight from it, and using information for better decision-making. It involves the process of collecting, preparing, managing, analyzing, explaining, and disseminating the data and analysis results. This is consistent with the current understanding of the broad scope of the field—see, for example, the CRA's use of the term.[17]

Note that the definition intentionally uses the term "data-based" rather than the more common "data-driven." The latter has frequently been interpreted as "data should be the main (only?) basis for decisions" since "data speaks for itself." This is wrong—data certainly holds facts and can reveal a story, but it only speaks through those who interpret it, and that can potentially introduce biases. Therefore, data should be one of the inputs to decision making, not the only one. Furthermore, "data-driven" has come to mean that it is possible to take data and analyze it by automated tools to generate automated actions. This is also problematic despite the recent popularity of, and over-reliance on, predictive and prescriptive analytics. Though data science has significant potential, and successful data-science applications are plentiful, there are sufficient misuses of

Figure 2. Data-science building blocks.



data to give us pause and concern: Google's algorithmic detection of influenza spread using social media data is one prominent example, while the Risk Needs Assessment Test used in the U.S. justice system is another. Therefore, "data-based" is the preferable phrase that signals that data-science deployments are aids to the decision maker, not decision makers themselves.[a]

There are significant differences between what was called data analysis (or analytics) and what the current understanding of data science entails.

Data-Science Ecosystem

Data science is inherently interdisciplinary: It builds on a core set of capabilities in data engineering, data analytics, data protection, and ethics—the four pillars of data science (see Figure 2). Some of this core is technical, some is not. Although the term "data science" is frequently used to refer only to data analysis, the scope is wider, and the contributing elements of the field should be properly recognized. The core is in close interaction with application domains that have the dual function of informing the appropriate technologies, tools, algorithms, and methodologies that should be useful to develop and leverage these capabilities to solve their problems. Data-science application deployments are highly sensitive to existing social and policy context, and these influence both the core technologies and the application deployments.

Data engineering. Data is at the core of data science; the type of data that is used is commonly referred to as big data. There is no universal definition of big data; it is usually characterized as data that is large (volume); multi-modal (variety) with many types of data: structured, text, images, video, and others; sometimes streaming at high speed (velocity); and has quality issues (veracity). These are known as the "four Vs" and addressing them appropriately is the domain of big-data management.[31] Data engineering in data science addresses two main concerns: the *management of big data* (including the computing platforms for its processing) and the *preparation of data* for analysis.

Managing big data is challenging but critical in data-science application development and deployment. These data characteristics are quite different than those that traditional data-management systems are designed for, requiring new systems, methodologies, and approaches. What is needed is a data-management platform that provides appropriate functionality and interfaces for conducting data analysis, executing declarative queries, and enabling sophisticated search. This exceeds the current state of the art, where individual systems are specialized toward a specific data model and require metamodels that allow for different levels of abstraction, and for more fluent integration and seamless interoperability. Data preparation is typically understood to be the process of dataset selection, data acquisition, data integration, and data-quality enforcement. Applying appropriate analysis to the integrated data will provide new insights that can improve organizational effectiveness and efficiency and result in evidence-informed policies. However, for this analysis to yield meaningful results, the input data must be appropriately prepared and trustworthy. The quality of the analysis model makes little difference; if the input data is not clean and trustworthy, then the results will not very valuable. The adage of "garbage in, garbage out" is real in data science. Data quality is an essential element in making the data trustworthy for analysis. It addresses the veracity characteristic of big data. Data quality is considered mission critical in data-science success and constitutes a major portion of the data-preparation effort for most organizations.

An important vehicle of data quality and data trustworthiness is metadata and metadata management. One particularly important metadata that deserves mention is provenance, which tracks the source of the original data. Another challenge is developing and instituting the appropriate system and tool support for managing provenance, and tracking data as it goes through the processing pipeline.

A very important aspect of data quality is data cleaning.[21] When data from multiple sources is used, there are bound to be inconsistencies, errors in data, and missing informa-

a Other suitable terms that have been used are "data-enhanced" or "data-enabled".

tion that must be corrected (cleaned). Techniques and methodologies for data cleaning are an important part of data engineering.

Data analytics. Data analytics is the application of statistical and ML techniques to draw insights from data under study and to make predictions about the behavior of the system under study. A first-level distinction in data analysis is made between inference and prediction. Inference is based on building a model that describes a system behavior by representing the input variables and relationships among them. Prediction goes further and identifies the courses of action that might yield the "best" outcomes. This classification can be made more finely grained by identifying four different classes: *descriptive*, which retrospectively looks at the historical data to answer the questions "What happened?" or "What does the data tell us?; *diagnostic*, which is also retrospective but goes beyond descriptive to answer the question "Why has that happened?"; *predictive*, a forward-looking analysis of historical data that provides calculated predictions of what is likely to happen; and *prescriptive*, which goes further by recommending courses of action. Predictive and prescriptive analytics together are usually called advanced analytics. The relationship among these is usually evaluated along two dimensions: complexity and value.[27] Going from descriptive to prescriptive, analysis becomes far more complex, but the value derived from it also substantially increases.

There are six data-analysis tasks (methods) commonly used in data science:[15,24] *clustering*, which finds meaningful groups or collections of data based on the "similarity" of data points (data points in the same cluster are more similar to each other than they are to data points in other clusters); *outlier detection*, which refers to identification of rare data items in a dataset that differ significantly from the majority of the data; *association rule learning* discovers interesting relationships between variables in a large dataset; *classification* finds a function (model) that places a given data item in one of a set of predefined classes; *regression* finds the function that relates one or more independent variables to a dependent variable; and *summarization* creates a more compact representation of the dataset.

As discussed earlier, an important data source in data science is streaming data. In this case, real-time analytics must be considered as data flows continuously. Real-time analytics is particularly difficult given that most analysis algorithms are computationally heavy and usually require multiple passes over the dataset, which is challenging in streaming data.

In a data-science project, an important consideration is the selection of appropriate techniques for the task and how they can be leveraged. Given the societal impact of data-science applications and deployments, the explainability of the analysis results is equally important.

Data protection. Data science's reliance on large volumes of varied data from many sources raises important data-protection concerns. The scale, diversity, and interconnectedness of data (for example, in online social networks) requires revisiting data-protection techniques that have been mostly developed for corporate data.[7,29]

It is customary to discuss the relevant issues under the complementary topics *data security* and *data privacy*. The former protects information from unauthorized access or malicious attacks, while the latter focuses on the rights of users and groups over data about themselves. Data security typically deals with data confidentiality, access control, infrastructure security, and system monitoring, and uses technologies such as encryption, trusted execution environments, and monitoring tools. Data privacy, on the other hand, deals with privacy policies and regulations, data retention and deletion policies, data-subject access requirement (DSAR) policies, management of data use by third parties, and user consent. Data privacy normally involves privacy-enhancing technologies (PETs). Although research on these topics is usually isolated, it is helpful to take a holistic and broader view, hence the term data protection is more appropriate and informative.

The characteristics of data used in data science pose unique challenges. Data volumes make the enforcement of access-control mechanisms more difficult and the detection of malicious data and use more challenging. The numerosity and variety of data sources make it possible to inject mis/disinformation, skewing the analysis results.[7] Data-science platforms are, by necessity, scale-out systems that increase the possibility of infrastructure attacks. These environments also increase the potential for surveillance. The variability and potentially high numbers of end users, and in many data-science deployments, the need for openness for sharing analysis results and for bolstering the analysis, opens the possibility of data breaches and misuse. These factors seriously increase the threats and the attack surface. Therefore, protection is required for the entirety of the data-science life cycle, from data acquisition to the dissemination of results, as well as for secure archiving or deletion. An implicit goal of data science is to gain access to as much data as possible, which directly conflicts with the fundamental least-privilege security principle of providing access to as few resources as necessary. Closing this gap includes careful redesign and advancement of security technologies to preserve the integrity of scientific results, data privacy, and to comply with regulations and agreements governing data access. Techniques that have been developed for privacy-preserving data mining are examples of this consideration.

Data-science ethics. The fourth building block of data science is ethics. In many discussions, ethics is bundled with a discussion of data privacy. The two topics certainly have a strong relationship, but they should be considered separate pillars of the data-science core. Literature typically refers to *data ethics* as ". . . the branch of ethics that studies and evaluates moral problems related to data, . . . algorithms, . . . and corresponding practices, in order to formulate and support morally good solutions."[16] The definition recognizes the three dimensions of the issue—data, algorithms, and practice.

▸ The *ethics of data* refers to the eth-

ical problems posed by the collection and analysis of large datasets and on issues arising from the use of big data in a diverse set of applications.

▸ The *ethics of algorithms* addresses concerns arising from the increasing complexity and autonomy of algorithms, their fairness, bias, equity, validity, and reliability.[18]

▸ The *ethics of practices* addresses questions concerning the responsibilities and liabilities of people and organizations in charge of data processes, strategies, and policies. The growing research in AI ethics tackles many of these issues.

Perhaps one of the most important concepts in data-science ethics is informed consent. Participants of data-science projects should have full information about the project, its objectives, and scope, and they should freely agree to participate. If data about participants is collected, they should have full knowledge of what is being collected and how it will be used (including by third parties) so they can agree to its collection and use.

One important issue in data-science ethics that has received significant attention is bias. *Oxford English Dictionary* defines bias as the "inclination or prejudice for or against one person or group, especially in a way considered to be unfair." Bias is inherent in human activities and decision making, and human biases are reflected in data science as biases in data and biases in algorithms. Bias in data can be introduced through what is included in the historical data used by the algorithms—for example, arrest records in the U.S. include more marginalized communities, primarily because they are over-policed. Bias in data may also be introduced due to under-representation—for example, data used in face-recognition systems comprises 80% whites, of which three-quarters are males. Algorithmic bias can occur due to the inclusion or omission of features in the algorithm/model. In ML deployments, this can occur during feature engineering. These features include individual attributes such as race, religion, and gender. The use of proxy metrics (for instance, using standardized examination scores to predict student success) can also lead to bias.

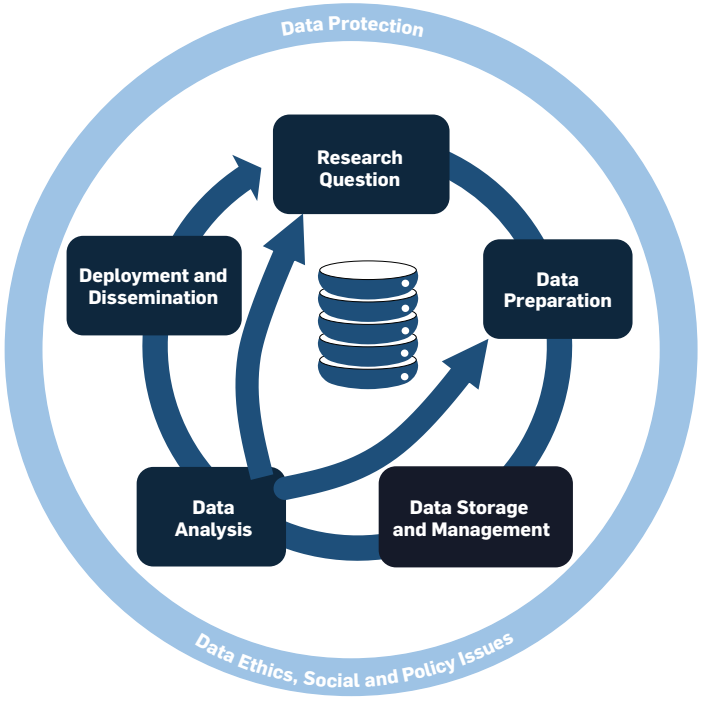
Although considerable attention has focused on the problem of bias, data-science ethics should be considered more generally consistent with the previously offered, broader definition of ethics. Some of the broader ethical considerations have overlapping concerns with data protection. Related to ethics of practice, societal norms that usually get coded in legislation are important (more on this in the next section). Consequently, while some of the ethical concerns are universal, others can be specific to a jurisdiction.

Broader ethical concerns also include: ownership of data; transparency, which refers to subjects knowing what data is collected about them and how it will be stored and processed (including informed consent by the subject); privacy of personal data, in particular the revealing of personal identifiable information; and intention regarding how data will be used, especially for secondary use.

Social and policy context. As noted earlier, data-science deployments are highly sensitive to the societal and policy contexts in which they are deployed. For example, what can be done with data differs in different jurisdictions. The context can be legal, establishing legal norms for data-science deployments, or it can be societal, identifying what is socially acceptable. Furthermore, there are significant intersections between social science and humanities and the core issues in data science. There are four central concerns: ownership, representation, regulation, and public policy. Obviously, there is overlap between these and the data-ethics concerns previously discussed.

Ownership. Data ownership, access, and use—particularly in terms of how individual data is generated, who owns and can access to it, and who profits from it—is a critical concern. At the societal and organizational levels, researchers analyze how economic systems are increasingly data-dependent in terms of both operations and revenue streams, and how pressures to collect and share ever-more intimate data may conflict with users' own calls for privacy and autonomy. Data privacy from a technical perspective was previously discussed, but it obviously has a signifi-

Figure 3. Data-science life cycle.



cant legal and social dimension that requires careful study (for example, Solove[35]).

Representation. A primary concern in the development of data-science technologies is ensuring diverse and equitable representation at all stages of the life cycle. This includes the evaluation of the training, tools, and techniques used in data science, including who designs them, who has access to them, and who is represented by them. Data representativeness is tightly tied to questions of marginalization and bias that appear throughout the design, data collection, analysis, and implementation of these technologies—for example, Richardson et al.[32] Another concern is how data is increasingly used to "speak" for users, often without their knowledge—changing individuals' relationships with their local communities, corporations, and state.

Regulation and accountability. Components of ethical data science also include a commitment to transparency and explainability in terms of analyzing the inputs of data-driven decision making, the algorithms applied, and determining how they lead to specific outputs and recommendations. Ensuring that the benefits and opportunities afforded by advances in data science equally benefit broader society involves accountability and regulatory practices at each stage of the data-science life cycle and are not only centered on laws and policy interventions, such as the General Data Protection Regulation (GDPR) and the Canadian Personal Information Protection and Electronic Documents Act (PIPEDA). They must also include efforts to include values-in-design, interventions for more accessible and inclusive design, and tools for ethical thought at the levels of training, education, and ongoing daily practice.

Public policy. There is a critical and urgent need to integrate data science into the analysis of public policy.[36] In an age where every Facebook, Twitter, and Instagram post is a data observation that can be archived and can become part of a historical dataset that can inform public policy, governments have been left behind in their ability to collect, aggregate, and analyze this data. With the necessary tools, this data could be managed and analyzed in a way that is explainable and that can be disseminated in a meaningful way to provide key insights. In a similar vein, there is a paradox in the large amount of "open" data that remains unused, along with concerns of a data deficit in terms of more information that could and should be collected to inform public policy.

Who "owns" data science is a topic of some discussion, primarily between statisticians and computer scientists.

Data-Science Life Cycle

The definition of data science previously given clearly identifies the process view of data science, namely that it consists of several processing stages, starting from data ingestion and eventually leads to better decisions, insights, and action. That process is called the data-science life cycle. Literature refers to the data life cycle, focusing only on data processing. A good definition of the data life cycle is given by the U.S. National Science Foundation Working Group on the Emergence of Data Science,[6] which identifies five linear stages: acquiring the data, cleaning it and preparing it for analysis, using the data through analysis, publishing the data and the methods used to analyze the data, and preserving/destroying the data according to policy. Variations of this data-life cycle model emerge in various proposals, some predating the above formulation—for example, Agrawal et al.,[2] Jagadish,[22] and Stodden.[37]

This life cycle model and its variants give the impression that the entire process is linear and unidirectional. Real project development hardly works in a linear fashion. An alternative model that is more iterative with built-in feedback loops has been proposed in the CRoss-Industry Standard Process for Data Mining (CRISP-DM) model for data-mining projects.[34] CRISP places data at the center and specifies a cyclical life cycle that is iterative and may be repeated over the lifetime of the project. PPDAC[26] is similar to CRISP-DM for statistical analysis tasks. Microsoft Team Data Science Access Life cycle[39] also emphasizes the iterative nature of the process.

The data-science life cycle proposed in this article (see Figure 3) derives

from and is built on those iterative models. It starts with the specification of the research question that may come from a particular application or may be an exploratory question. A good understanding of the research question is important, since it normally drives the entire process. The next step is data preparation, which includes determining which datasets are needed and available; selecting the appropriate datasets from within this larger set; ingesting the data; and addressing data-quality issues, including data cleaning and data provenance. The third step is the proper storage and management of the data, including big-data management. Specifically, data needs to be integrated, decisions need to be made about the storage structures for data for efficient access, appropriate storage structures need to be chosen and designed, suitable access interfaces must be specified, and provisions need to be made for metadata management, in particular for provenance data. The prepared and suitably stored data is then open for analysis. In particular, the appropriate statistical and ML model(s) is/are selected/developed, feature engineering is performed to identify the most appropriate model parameters, and model validation studies are conducted to determine the model's suitability. If model validation is successful, the next step is deployment and dissemination, which involves different activities depending on the particular project and application. In some cases, the analysis and processing of data needs to be performed on a continuing basis, so deployment involves maintaining and monitoring the system over time. In other cases, deployment may involve compilation and dissemination of analysis results and their explanation. Dissemination of the analysis results, and sometimes even the curated data, is an important aspect of this phase. Open data, which is data that "anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness"[b]) is an important part of dissemination. Many governments and private institutions are adopting Open Data Principles stating that data should not only be open but should be complete, accurate, primary, and released in a timely manner. These properties make this data very valuable to data scientists, journalists, and the public. When open data is used effectively, data scientists can explore and analyze public resources, which allows them to question public policy, create new knowledge and services, and discover new value for social, scientific, or business initiatives.

Data science should be viewed as a unifying force that connects several fields, some of which are STEM and some are not.

Problems during the analysis phase may result in the process returning to either reformulating the research question (it may be underspecified or overspecified, making model building infeasible) or cycling back to data preparation if the model requires other or different data that has not been prepared. As noted earlier, data-science deployments are not "one-and-done." Following deployment, there must be constant monitoring—perhaps the environment changes, the data changes, or there is a deeper understanding of the research question that results in its revision and improvement. Thus, the process cycles as a dialectic process—every time the process comes back to the research question, we are at an elevated understanding of what needs to be studied. It is important to recognize that the stages in the life cycle are not isolated; the boundaries between stages are fuzzy, and there are important and interesting issues that arise at their intersections.

There is a continuous bi-directional interaction between this life cycle and the data-protection issues. Similarly, the ethical concerns, social norms, and policy framework impact each of the phases, sometimes even preventing the initiation of specific data-science studies.

Data Science Is Interdisciplinary

Who "owns" data science is a topic of some discussion, primarily between statisticians and computer scientists. This discussion bleeds into the question of who data scientists are, which then leads to different educational models of data science. Given the centrality of data to both disciplines, this discussion is perhaps not

b See http://opendefinition.org.

surprising. The concern among statisticians regarding the field of data science is long-standing. Given the early promotion of data analytics as an important topic by Tukey, there is a strong feeling among statisticians that they own (or should own) the topic. In a 2013 opinion piece, Davidian[13] laments the absence of statisticians in a multi-institution data-science initiative and asks if data science is not what statisticians do. She indicates that data science is "described as a blend of computer science, mathematics, data visualization, machine learning, distributed data management—and statistics," almost disappointed that these disciplines are involved along with statistics. Similarly, Donoho laments the current popular interest in data science, indicating that most statisticians view new data science programs as "cultural appropriation."[14]

There is a well-known argument put forth by Conway[10] on the nature of data science. He proposes three main areas organized as a Venn diagram: hacking skills, mathematics and statistics, and substantive experience. The hacking skills he contends to be important are the ability "to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically." Mathematics and statistics knowledge, at the level of "knowing what an ordinary least squares regression is and how to interpret it" is required to analyze data. The substantive experience is about the research problem that may come from an application domain or a specific research project. The Conway diagram, as it has come to be known, has become popular in these ownership debates by those who do not see a central role for computer science in data science, because Conway argues that hacking skills have nothing to do with computer science: "This, however, does not require a background in computer science—in fact, many of the most impressive hackers I have met never took a single CS course."

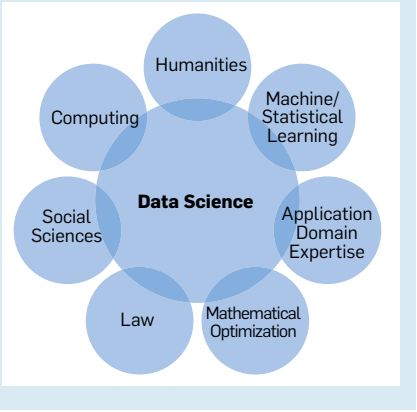
The computer science view espouses instead the centrality of computing. One such view has been championed by Ullman,[42] who also uses a Venn diagram to express his viewpoint and counters Conway, since he considers "algorithms and techniques for processing large-scale data efficiently as the center of data science." Ullman claims that the two big knowledge bases of data science are computer science and domain science (that is, the application domain), and their intersection is where data science resides. He sees, rightly, ML as part of computer science. He argues, again rightly, that some of ML is used for data science, but there are ML applications that are outside of data science. His diagram shows that there are aspects of data science that require computer-science techniques which have nothing to do with machine learning—data engineering, as previously discussed, would fall in that category. I suspect that some of these points may not be controversial. Where the argument is likely to be challenged is that, in his view, mathematics and statistics "do not really impact domain sciences directly" albeit their importance in computer science. Within computer science, there is the discussion regarding the relationship of AI/ML with data science, with some indicating that data science is part of AI, which was addressed in this article's first section.

A more balanced view has been put forth by Marina Vogt,[c] who indicates data science as sitting at the intersection of computer science, mathematics and statistics, and domain knowledge. This is more in line with the view put forward by ACM in its curriculum proposal: "Data science is an interdisciplinary endeavor between computer science, mathematics, statistics, and applied areas such as natural sciences."[1] However, even this viewpoint is very STEM-centric and leaves out many topics that are of interest to data science as a field.

These discussions and the resulting controversies are not helpful or necessary; they do not move the data-science agenda forward. No single community "owns" data science—it is too big, the challenges are too great, and it requires involvement

c The original article can no longer be found, but Vogt's viewpoint can be found here: http://www.policyhub.net/node/212.

Figure 4. Unifying view of data science.



from many disciplines. Creation and use of knowledge are fundamental human activities that span millennia. This activity is at the core of what we can define as being our collective human culture. Attempts to splinter the core of human achievement through an attribute of ownership is, at its best, a narrow parochial view, and, at its worst, ascendancy of self-interest and greed.

Data science should be viewed as a unifying force that connects several fields (see Figure 4), some of which are STEM and some are not. I go back to my discussion about stakeholders, who are diverse. Ownership arguments take place within one stakeholder group—STEM people who focus on foundational techniques and the underlying principles. Within this group, it is important to recognize and accept that there are communities with complementary and sometimes overlapping interests: computer scientists who bring expertise in computational techniques/tools that can effectively deal with scale and heterogeneity, statisticians who focus on statistical modeling for analysis, and mathematicians who have much to contribute with discrete and continuous optimization techniques and precise modeling of processes. However, this is only one stakeholder group; I have identified two others. One danger in such a unifying view is not finding the right balance between inclusiveness in accepting the contributions of all these fields and identifying the core of data science. I believe arguments made earlier in this article have established the core, so this danger is averted.

Conclusion

Despite its recent popularity, the field of data science is still in its infancy and much work needs to be done to scope and position it properly. The success of early data-science applications is evident—from health sciences, where social-network analytics have enabled the tracking of epidemics; to financial systems, where guidance of investment decisions is based on the analysis of large volumes of data; to the customer care industry, where advances in speech recognition have led to the development of chatbots for customer service. However, these advances only hint at what is possible; the full impact of data science has yet to be realized. Significant improvements are required in fundamental aspects of data science and in the development of integrated processes that turn data into insight. Current developments tend to be isolated to subfields of data science and do not consider the entire scoping as discussed in this article. This siloing is significantly impeding large-scale advances. As a result, the capacity for data-science applications to incorporate new foundational technologies is lagging.

The objective of this article is to lay out a systematic view of the data-science field and to reinforce the key takeaways: It is important to clearly establish a consistent and inclusive view the entire field; it is essential to define the core of the field while being inclusive to avoid becoming a catch-all or whatever the particular circumstances allow; it is critical to take a holistic view of activities that comprise data science; and a framework needs to be established to facilitate cooperation and collaboration among a number of disciplines.

Acknowledgments

A preliminary version of the ideas in this article appeared in an opinion piece in a 2020 bulletin of the *IEEE Technical Committee on Data Engineering 43*, (3) 3–11. My views on data science were sharpened in discussions with many colleagues as we worked on several data-science proposals. I thank colleagues (too many to list individually) who participated in these initiatives and taught me different aspects of the issues; they will see their fingerprints in the text of this article. I would especially like to acknowledge the many early discussions on framing the field and the relevant joint work with Raymond Ng and Nancy Reid. I thank Samer Al-Kiswani, Angela Bonifati, Khuzaima Daudjee, Maura Grossman, John Hirdes, Florian Kerschbaum, Jatin Matwani, Renée Miller, and Patrick Valduriez for feedback on all or part of this article. I very much appreciate the feedback from the anonymous reviewers, who pointed to weaknesses in some of my arguments and challenged me to clarify others. These helped improve the article.

References

1. ACM Data Science Task Force. Computing competencies for undergraduate data science curricula. Association for Computing Machinery (January 2021); https://bit.ly/3Vs7z3M.
2. Agrawal, D. et al. Challenges and opportunities with big data. The Computing Community Consortium of the CRA (2012); http://bit.ly/3UG0EDN
3. Ahalt, S.C. Why data science? The National Consortium for Data Science (October 2013); http://bit.ly/3KwC0AV.
4. Alvargonzález, D. Multidisciplinarity, interdisciplinarity, transdisciplinarity, and the sciences. *Intern. Studies in the Philosophy of Science 25*, 4 (2011), 387–403; http://bit.ly/3zRMBBD.
5. A new paradigm for business of data. World Economic Forum; https://bit.ly/3VupVBf.
6. Berman, F. et al. Realizing the potential of data science. *Communications of the ACM 61*, 4 (2018), 67–72; http://bit.ly/3Uvu8nS.
7. Bertino, E. and Ferrari, E. Big data security and privacy. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, S. Flesca, S. Greco, E. Masciari, and D. Saccà (eds.). Springer International Publishing (2018), 425–439; https://bit.ly/3obRJhh.
8. Chakravaram, V. et al. The role of big data, data science and data analytics in financial engineering. In *Proceedings of the 2019 Intern. Conf. on Big Data Engineering*, Association for Computing Machinery, 44–50; http://bit.ly/3oawNas.
9. Choi, B.C.K and Pak, A.W.P. Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. *Clinical and Investigative Medicine 29*, 6 (2006), 351–364.
10. Conway, D. The data science Venn diagram. *DrewConway.com* (2015); http://bit.ly/3mtgtkF.
11. *Data Science Applied to Sustainability Analysis*. J. Dunn and P. Balaprakash (eds.), Elsevier (2021); http://bit.ly/414PQSd.
12. *Data Science for Healthcare*, S. Consoli, D.R. Recupero, and M. Petković (eds.), Springer (2019); http://bit.ly/3MY7Old.
13. Davidian, M. Aren't we data science? *Magazine of American Statistical Association* (2013); http://bit.ly/413AwVH.
14. Donoho, D. 50 years of data science. *J. Computational and Graphical Statistics 26*, 4 (2017), 745–766; http://bit.ly/3zU2qrh.
15. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. *AI Magazine 17*, 3 (March 1996), 37–54; https://bit.ly/4233xSa.
16. Floridi, L. and Taddeo, M. What is data ethics? *Philosophical Trans. of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*, 2083 (2016); https://bit.ly/3LxA5MS.
17. Getoor, L. et al. Computing research and the emerging field of data science. Computing Research Association (2016); https://bit.ly/3Vwz7oH.
18. Grimm, P.W., Grossman, M.R., and Cormack, G.V. Artificial intelligence as evidence. *J. Technology and Intellectual Property 9*, 1 (2021), Article 2. https://bit.ly/3nMPYax.
19. Hey, T., Tansley, S., and Tolle, K. The fourth paradigm: Data-intensive scientific discovery. Microsoft Research (October 2009); https://bit.ly/3nnpxYE.
20. *IFIP Guide to Concepts and Terms in Data Processing Volume 1*. Intern. Federation for Information Processing, North-Holland Publishing Company (1971); https://bit.ly/3LObrZS.
21. Ilyas, I.F. and Chu, X. *Data Cleaning*. Association for Computing Machinery, New York, NY, USA (2019); https://bit.ly/3LMnaYN.
22. Jagadish, H.V. Big data and science: Myths and reality. *Big Data Research 2*, 2 (2015), 49–52; https://bit.ly/44lxEFL.
23. Johnson, G. The world: In silica fertilization; All science is computer science. *The New York Times* (March 25, 2001).
24. Kelleher, J.D. and Tierney, B. *Data Science*. MIT Press, Cambridge, Mass. (2018).
25. *Machine Learning and Data Science in the Power Generation Industry*. P. Bangert (ed.), Elsevier (2021); http://bit.ly/4174sAg.
26. MacKay, R.J. and Oldford, R.W. 2000. Scientific method, statistical method and the speed of light. *Statistical Science 15*, 3 (2000), 254–278; https://bit.ly/41W5m3e.
27. Maydon, T. *The 4 Types of Data Analytics* (2017); https://bit.ly/42jIeLA.
28. Milligan, I. *History in the Age of Abundance? How the Web is Transforming Historical Research*. McGill University Press, Montreal (2019).
29. Moura, J. and Serrão, C. Security and privacy issues of big data. In *Cloud Security: Concepts, Methodologies, Tools, and Applications*, IGI Global (April 2019), 1598–1630; https://bit.ly/3LvfLM6.
30. Naur, P. *Concise Survey of Computer Methods*. Petrocelli Books (1974); https://bit.ly/420GdEs.
31. Özsu, M.T. and Valduriez, P. *Principles of Distributed Database Systems (4th Edition)*, Springer (2020).
32. Richardson, R., Schultz, J.M., and Crawford, K. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online 95*, 15 (2019), 15–55; https://bit.ly/3NAdeTx.
33. Sarker, I.H. Smart city data science: Towards data-driven smart cities with open research issues. *Internet of Things 19* (2022), 100528; https://bit.ly/40S0I4Q.
34. Shearer, C. The CRSIP-DM model: The new blueprint for data mining. *J. Data Warehousing 5* (2000), 13–22.
35. Solove, D.J. A taxonomy of privacy. *University of Pennsylvania Law Review 154*, 3 (2006), 477–560.
36. Steif, K. *Public Policy Analytics: Code and Context for Data Science in Government*. CRC Press (2021); https://bit.ly/413JgKB.
37. Stodden, V. The data science life cycle: A disciplined approach to advancing data science as a science. *Communications of the ACM 63*, 7 (2020), 58—66; https://bit.ly/42j3DEY.
38. Supriya, P. et al. Trends and application of data science in bioinformatics. In *Trends of Data Science and Applications: Theory and Practices*, S.S. Rautaray, P. Pemmaraju, and H. Mohanty (eds.), Springer Singapore (2021), 227–244; https://bit.ly/42cpiym.
39. Tabladillo, M. et al. The team data science process life cycle. Microsoft Learn (2022); https://bit.ly/4107xBt.
40. The world's most valuable resource is no longer oil, but data. *The Economist* (May 2017).
41. Tukey, J.W. The future of data analytics. *The Annals of Mathematical Statistics 33* (1962), 1–67.
42. Ullman, J.D. The battle for data science. *IEEE Data Engineering Bulletin 43*, 2 (2020), 8–14; https://bit.ly/3p3RzJr.

M. Tamer Özsu (tamer.ozsu@uwaterloo.ca) is a professor at the Cheriton School of Computer Science, University of Waterloo, Ontario, Canada.

Watch the author discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/data-science-systematic-treatment

ACM BOOKS

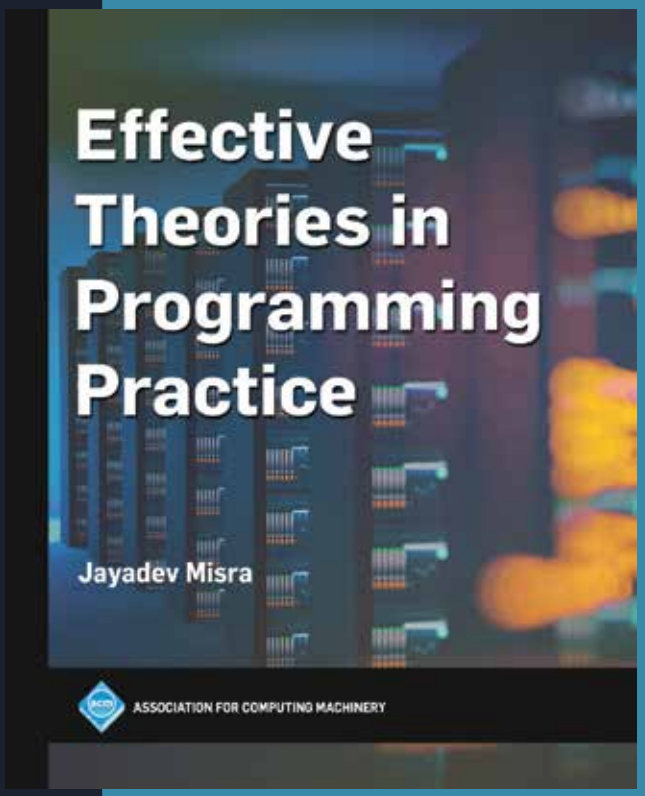
Collection II

Set theory, logic, discrete mathematics, and fundamental algorithms (along with their correctness and complexity analysis) will always remain useful for computing professionals and need to be understood by students who want to succeed. This textbook explains a number of those fundamental algorithms to programming students in a concise, yet precise, manner. The book includes the background material needed to understand the explanations and to develop such explanations for other algorithms. The author demonstrates that clarity and simplicity are achieved not by avoiding formalism, but by using it properly.

The book is self-contained, assuming only a background in high school mathematics and elementary program writing skills. It does not assume familiarity with any specific programming language. Starting with basic concepts of sets, functions, relations, logic, and proof techniques including induction, the necessary mathematical framework for reasoning about the correctness, termination and efficiency of programs is introduced with examples at each stage. The book contains the systematic development, from appropriate theories, of a variety of fundamental algorithms related to search, sorting, matching, graph-related problems, recursive programming methodology and dynamic programming techniques, culminating in parallel recursive structures.

ABOUT THE AUTHOR
Jayadev Misra is the Schlumberger Centennial Chair Emeritus in computer science and a University Distinguished Teaching Professor Emeritus at the University of Texas at Austin. Professionally he is known for his contributions to the formal aspects of concurrent programming and for jointly spearheading, with Sir Tony Hoare, the project on Verified Software Initiative (VSI).

http://books.acm.org



Effective Theories in Programming Practice

Jayadev Misra
University of Texas at Austin

ISBN: 9781450399715
DOI: 10.1145/3568325

DOI:10.1145/3571723

A case study reveals the theoretical analysis of algorithms is not always as helpful as standard dogma might suggest.

BY PAUL MEDVEDEV

Theoretical Analysis of Sequencing Bioinformatics Algorithms and Beyond

WHEN I ASK first-year computer science undergraduate students how to quantify the speed of an algorithm, they look at me puzzled and tell me to just run the algorithm and see how long it takes. Such empirical analysis, in fact, is the most direct and natural way to measure algorithm performance. However, it has long been understood to have many shortcomings.[45] To overcome these shortcomings, computer scientists developed ways to theoretically analyze algorithm performance. The most common technique for this is traditional worst-case analysis. For example, we say merge sort runs in $O(n \log n)$ worst-case time, which formally means that there exists a constant c such that for any large-enough input of n elements, merge sort takes at most $cn \log n$ time. Other more sophisticated techniques, such as parametrized analysis, average-case analysis, or semi-random models, better capture the properties of real data.[44] Additionally, theoretical analysis can be used to measure not just speed but other aspects of algorithm performance, such as memory usage or accuracy. When undergraduate students take an algorithms course, they finally learn about the theoretical analysis of algorithms and how to use it to capture general patterns of performance that empirical analysis does not.

The most direct impact of such theoretical analysis is in applied algorithms, that is, algorithms implemented and applied to real data (in contrast to complexity theory, where such analysis serves the purpose of understanding the hierarchy of problem, rather algorithm, complexity). The goals of the theoretical analysis of applied algorithms, denoted here as TA3, are twofold:[44] One goal is to *predict* the empirical performance of an algorithm, either in an absolute sense or relative to others. The second goal is to be a yardstick that drives the design of novel algorithms that perform well in practice. TA3 has achieved its goals with resounding success, being directly responsible for the design and performance prediction of many algorithms used in practice (for example, Dijkstra's shortest path algorithm). Because of this success, algorithms instructors typically jump into theoretical analysis with only a cursory justification of why it is needed. To put it bluntly, the fact that TA3 achieves the two goals has become a dogma of computer science.

However, fast-paced application domains pose a challenge to TA3. In this article, I use the field of sequencing bioinformatics (SeqBio) as a case study, where I posit that TA3 has failed to achieve its stated goals. SeqBio is an interdisciplinary field that uses algorithms to extract biological mean-

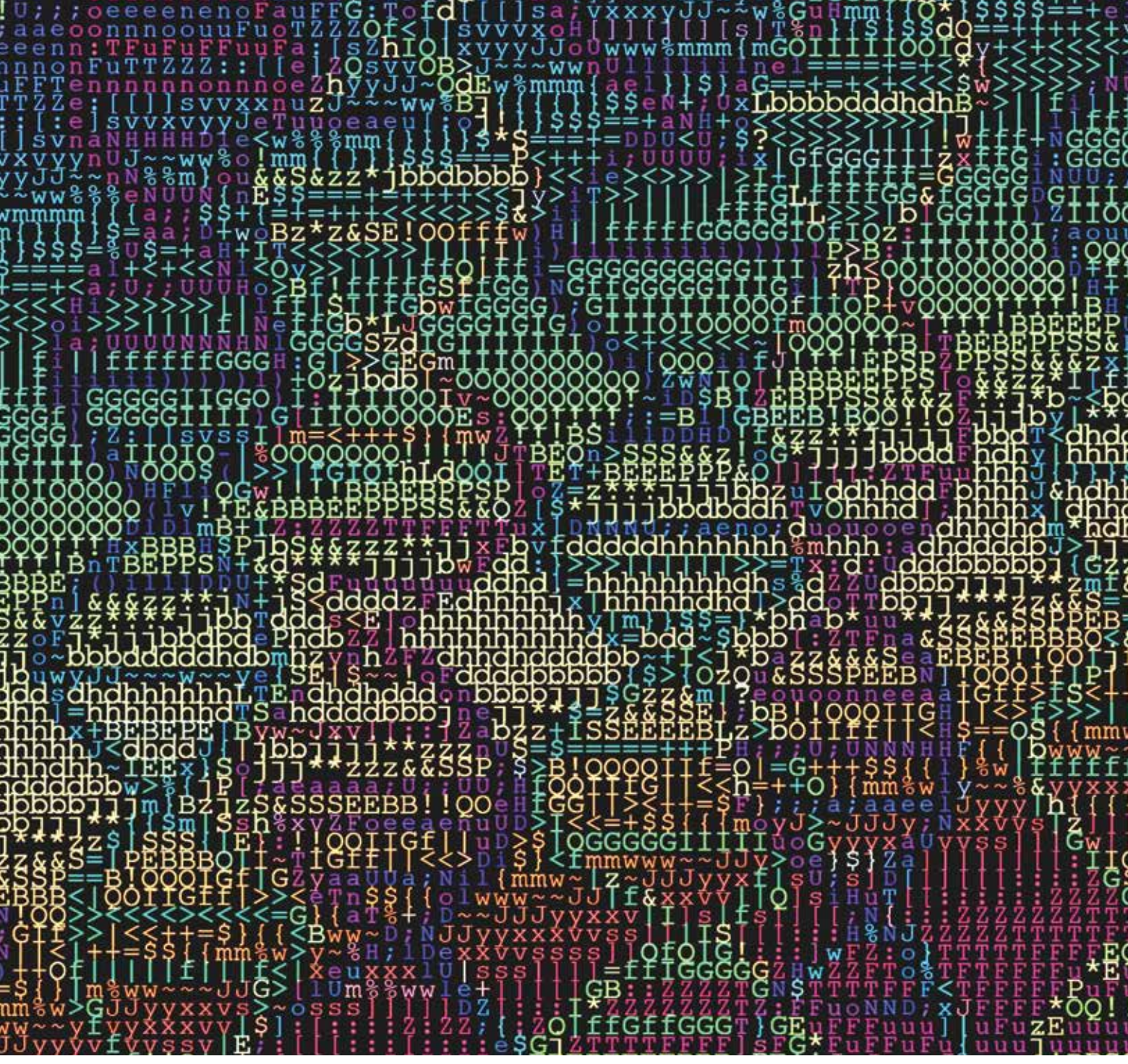


IMAGE BY 3D_VICKA

ing from sequencing data. SeqBio has revolutionized the life sciences, with algorithms developed by computer scientists (for example, Bankevich et al.[3] and Langmead et al.[23]) enabling projects such as the Earth Microbiome Project,[15] the Vertebrate Genomes Project,[42] and the Cancer Genome Atlas.[18] It is also in the process of revolutionized healthcare, enabling projects such as Obama's Precision Medicine Initiative. However, the vast majority of SeqBio papers either do not perform TA3 (for example, Bankevich[3]) or perform traditional worst-case analysis only to conclude that it is not a good predictor of the algorithm's performance in practice (for example, Marschall et al.[29]). I will demonstrate two concrete examples of TA3's failure in SeqBio: the problems of genome assembly and structural variation detection. I will also give one encouraging example of success: *k*-mer data structures and then catalog some of the challenges of applying theoretical analysis in SeqBio, argue why empirical analysis is not enough, and offer a vision for improving the relevance of theoretical analysis to SeqBio. By recognizing the problem, understanding its roots, and providing potential solutions, this work can hopefully be a crucial first step toward making TA3 more relevant in SeqBio and other fast-paced application domains.

Before proceeding, it is useful to make the distinction between direct and indirect influence of TA3 in SeqBio. Without question, TA3 can be credited with the development and analysis of methods which have become part of SeqBio's toolbox, for example, integer linear programming, clustering algorithms, sketching techniques, and machine learning. For the two problems we will investigate, combinatorial algorithms for optimization problems with

theoretical guarantees form the backbone of many tools[3,13,35,46] for assembly and for structural variant detection.[31] However, this article is concerned with direct influence, that is, situations where theoretical analysis was applied to problems specific to SeqBio, or at least for which SeqBio was a major motivation.

Accuracy of Genome Assemblers

SeqBio algorithms work with data generated by various instruments that repeatedly sample short substrings (called *reads*) from a long genome sequence.[28,a] The locations are chosen semi-uniformly at random but are not output by the instrument; the output only contains the string sequence of each read. Moreover, the reads may contain errors and hence not be exact substrings of the genome. Such a sequencing experiment can generate billions of reads at ever-decreasing costs. There are many applications of sequencing technology, but here I focus on the genome assembly problem, a classical SeqBio problem.

The genome assembly computational problem is to take a sequencing experiment from a single genome and to reconstruct the full DNA sequence of that genome.[47] There are dozens of widely used assemblers, with hundreds more at the prototype stage. Genome assembly algorithms have enabled genome-wide studies of thousands of species and have a biological impact that is difficult to overstate. The most important aspect of an assembler's performance is its accuracy, since the resources required to collect a DNA sample usually outweigh the computational resources of an assembler. In this section, we will describe the various attempts to apply theoretical analysis to predict the accuracy of assemblers and design more accurate assemblers.

Figure 1 illustrates an example of a simple assembly algorithm. However, several practical factors complicate this simple picture. First, reads can have sequencing errors which introduce erroneous vertices and edges into the assembly graph. Second, some parts of the genome are not covered by reads, introducing gaps in the graph. Third, repetitive sequences are prevalent and make the graph structure more convoluted.[21,33] These and other factors make it challenging to keep the output of the assembler accurate.

Theoretical analysis of assembler accuracy cannot be credited with the design of any of the widely used assemblers, nor has it led to any theoretical analysis that can predict the accuracy of an assembler on real data.

One of the earliest theoretical measures of accuracy is the likelihood of the reads given an assembly. The idea of building an assembler to maximize this likelihood was originally proposed in Myers[32] and later pursued in Boza et al.,[5] Howison et al.,[16] Medvedev et al.,[30] and Varma et al.[51] Much of the work centers on finding an appropriate likelihood function, that is, one which models the intricacies of the sequencing process. Unfortunately, these formulations have not directly led to any state-of-the-art assemblers, leaving the design goal of TA3 unfulfilled. The reasons for this are not clear from the literature, but I will discuss later.

I am not aware of any work that attempted to use likelihood to theoretically predict algorithm performance. Some works did explore the idea of using a likelihood score to evaluate the accuracy of an assembly.[12,14,17,24,40,52] Though these tools have been widely used, they are not designed to give a theoretical accuracy of an assembler but, rather, to be run on a concrete output. As such, they cannot predict in advance how an assembler will empirically perform, leaving the prediction goal of TA3 unfulfilled.

A later approach[7] is to evaluate an assembler by the conditions under which it can fully reconstruct the original genome sequence. Such conditions could, for example, be the number of reads needed or the highest error that could be tolerated. This accuracy framework did lead to the design of a new assembler called Shannon that has been used in practice,[19] thus partially fulfilling the design goal. However, these conditions rarely arise in practice (that is, the genome usually has too many repetitive sequences to be reconstructed completely and unambiguously). Therefore, it is not clear if designing algorithms to optimize this would lead to other assemblers that perform well in practice. In terms of the prediction goal, this framework was also applied to theoretically predict the accuracy of some simple assembly strategies;[7] however, it has not been applied to predict the accuracy of

a In recent years, the reads have become much longer; however, many assembly algorithms predate this advance.

any other assemblers used in practice. The common challenge to all prediction attempts such as this one is that most assembly algorithms rely heavily on ad hoc, hard-to-analyze heuristics.

One more recent approach is to measure what percentage of all the substrings that could possibly be inferred to exist in the genome are output by the assembler.[50] However, this framework has proven technically challenging to apply to real data, for example, to account for sequencing errors and gaps in coverage. It has not yet led to the design of a new assembler or to the accuracy prediction of assemblers, though work is ongoing.[8]

In summary, theoretical analysis of assembler accuracy cannot be credited with the design of any of the widely used assemblers, nor has it led to any theoretical analysis that can predict the accuracy of an assembler on real data. In practice, assemblers are designed heuristically to perform well on a set of empirically observed metrics,[34] such as the lengths of the segments that the assembler reports or the recovery of genes whose sequences are conserved across different species.[43] Additional validation is performed by measuring the agreement with data from an orthogonal sequencing technology, where the sequencing errors have different patterns.[41] Assembly algorithms are usually developed to perform well on publicly available datasets and often validated by the exact same datasets.

The shortcomings of TA3 have been felt in practice. The Assemblathon 2 competition[6] performed an empirical evaluation of assemblers and found the ranking of different assemblers according to their relative accuracy depended on the dataset, on the evaluation metrics being used, and on the parameter choices made in the evaluation scripts. These are exactly the shortcomings of empirical evaluation that TA3 is intended to address. Moreover, the assemblers themselves are simply not as good as they could be, or, as one of the reviewers concluded, "on any reasonably challenging genome, and with a reasonable amount of sequencing, assemblers neither perform all that well nor do they perform consistently."[49] In the last five years, the practical situation has to some extent improved due to newer sequencing technologies that generate higher quality data. Nevertheless, the experience of the Assemblathon 2 competition is instructive to appreciate the limitations of solely empirical analysis in SeqBio.

Accuracy of Structural Variation Detection Algorithms

Once a genome is assembled for a species, it forms what is called a reference genome. Follow-up studies then sequence different individuals of the same species but do not perform a de novo genome assembly. Instead, they catalog the variations between the sequenced genome and the reference, under the assumption that the genome is unchanged in places where there is no alternative evidence. Variants that affect large regions of more than 500 nucleotides, called structural variants, are responsible for much genomic diversity and are linked to numerous human diseases, including cancer and myriad neurodevelopmental disorders.[53] Algorithms for detecting structural variation started to appear around 2008 (see Mahmoud et al.[26] and Medvedev[31] for surveys) and a recent assessment identified at least 69 usable tools.[22] As with genome assembly, the most important aspect of algorithm performance is accuracy.

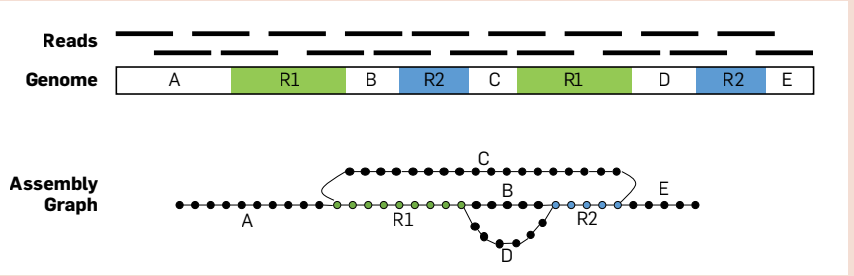
Figure 2 illustrates the types of signatures that algorithms can use to detect structural variants. What complicates the algorithm's task is that the same signature can sometimes be explained by alternate events, the signatures of multiple events can overlap, and repetitive sequences can make it difficult to find the correct location of a read.

Most tools are heuristics with no theoretical analysis of their accuracy. There are some exceptions, when a probabilistic formulation is used to achieve a desired false discovery rate (for example, Marschell[29]). In such cases, TA3 can take some credit for the design of the algorithm. However, accuracy greatly depends on the type of variant (for example, deletions are easier to detect than duplications) and on the location of the variant (for example, repetitive sequences make variants harder to detect). Thus, a useful analysis of accuracy requires a statistical model for the distribution of variant type, location, and relative frequency. But coming up with realistic models is challenging, as our understanding of the biological process that generates structural variants is limited. Therefore, even when accuracy is predicted theoretically, it does not correspond to what is observed in practice because the models are too idealized.[29] Thus, even in the limited cases where TA3 has been applied, it has not achieved its prediction goals.

As with genome assembly, the al-

Figure 1. Illustration of a simple assembly algorithm.

The figure shows a hypothetical genome composed of several long segments. Two segments are repeated: the green segment R1 appears twice, as does the blue segment R2. The other segments are assumed to be repeat-free, in the sense that all the substrings of length above a certain threshold k are unique. A potential set of sampled reads is shown above the genome, lined up according to where they come from in the genome. Their location is not known by the algorithm but is shown here for clarity. A simple assembly algorithm would construct an assembly graph from the reads, shown at the bottom. The nodes are all the k-mers (that is, substrings of length k) appearing in the reads and there is an edge between a pair of k-mers that follow one another in at least one read. The genome is a walk in this graph that covers all the vertices (A, R1, B, R2, C, R1, D, R2, E), however, there is more than one walk with this property (for example, A, R1, D, R2, C, R1, B, R2, E). A simple assembly algorithm would not risk making a mistake and would output only the walks in the graph that it is confident appear as sub-walks of any genome walk. In this case, the output could be the spelling of the seven walks A, B, C, D, E, R1, and R2.



gorithms used in practice suffer from many of the limitations that TA3 is intended to address. Algorithms are typically evaluated empirically, using simulated data or an established benchmark. Two recent studies assessing the empirical accuracy of algorithms[9,22] found the tools suffered from low recall and the ranking of the tools according to accuracy varied greatly across different subtypes of variants. Empirical evaluation is hampered by the same lack of models that hampers theoretical evaluation and Cameron et al.[9] warned developers against considering "simulation results representative of real-world performance." In fact, accuracy on simulated data is typically much higher for most tools then on real data.[22] As with genome assembly, the problems described in Cameron[9] are the types inherent to empirical-only evaluation: "But with newly published callers [algorithms] invariably reporting favourable performance, it is difficult to discern whether the results of these studies are representative of robust improvements or due to the choice of validation data, the other callers selected for comparison, or over-optimisation to specific benchmarks."

Memory Usage of *k*-Mer Data Structures

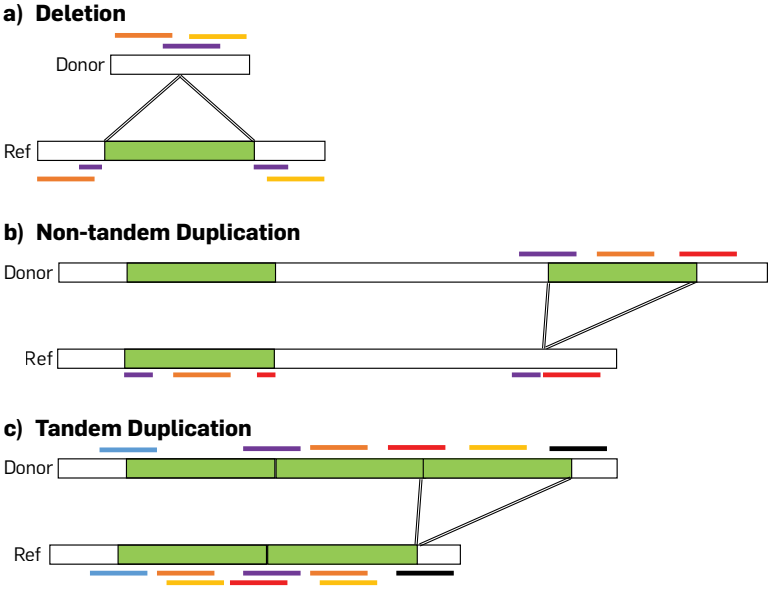
Sequencing data is often reduced to a collection of *k*-long strings (called *k*-mers) that are stored in a variety of data structures. For example, breaking reads into constituent *k*-mers is part of many assembly algorithms (see Figure 1). The exact data structure depends on the types of queries that need to be supported, the type of associated data maintained, and the source of the *k*-mer set. Examples of queries include simple membership queries (such as, is a *k*-mer present in the data structure?) and group membership queries (Does a given bag of *k*-mers have at least 70% of its *k*-mers present in the data structure?). Examples of associated data include count information (for example, how often does a *k*-mer occur in a set of reads?) or experiment information (Given multiple experiments, which experiments contain the *k*-mer?). Data structures to store *k*-mers have become ubiquitous in SeqBio and form the backbone of hundreds of tools (for surveys, see Chikhi et al.,[10] and Marchet et al.[27]). The theoretical analysis of their memory is a rare bright light in the theoretical analysis of SeqBio algorithms and I discuss it here to illustrate TA3's potential for success in SeqBio.

Many of the techniques used to analyze the memory used by *k*-mer data structures have been borrowed from the field of compact data structures.[36] The analysis of compact data structure memory differs from traditional worst-case analysis in that the higher-order terms are often written without asymptotic notation (for example, $4n + o(n)$ instead of $O(n)$). This helps distinguish algorithms whose memory differs by a constant factor, at the expense of a more technically involved analysis.

Using this type of analysis as a yardstick has led to the design of several *k*-mer data structures that perform well on real data and are included in broadly used software.[1,2,4,11,37,48] In one example, an analysis led to the design of a compact representation of a popular *k*-mer data structure called the *de Bruijn graph* that uses $4n + o(n)$ bits, where n is the number of edges.[4] This data structure uses very little memory in practice[39] and forms the core of the widely used MEGAHIT assembler.[25] Another successful example of a *k*-mer data structure is the pufferfish index,[2] which forms part of the popular Salmon[38] software. In addition to satisfying the design goal, this type of analysis has been used to correctly predict how much memory *k*-mer data structures will use in practice and compare the relative performance of different methods (for example, quantifying the trade-offs between memory and query time of various indices).

Figure 2. Illustration of structural variation detection.

Each panel shows the sequenced donor genome (top rectangle) and the species reference genome (bottom rectangle). It shows the reads above the donor genome according to their origin position; it shows the reads below the reference according to where a string pattern matching algorithm would place them (in more technical terms, an alignment). In panel (a), the donor genome has the green region deleted; in panel (b), the donor has the green region duplicated and inserted far away in the genome; in panel (c), the reference has two copies of the green region while the donor has three. Observe how each event leaves behind a signature that can be detected by an algorithm. In (a), the purple read is split into two partial matches, the position of each indicating the boundary of the deletion. In (b), the purple and red reads each have two partial placements, with one end at the insertion location and the other at the edge of the duplicated sequence. In panel (c), all the reads have normal-looking placements; however, there are more reads mapping to the green region then one would expect if there was no duplication.



Thus, theoretically analyzing memory by retaining the constant for the higher-order terms has satisfied both the design and prediction goals. Such a TA3 success, though rare in SeqBio, demonstrates it is indeed possible for TA3 to have an impact in SeqBio.

Empirical Analysis Is Not Enough

Despite the limitations of theoretical analysis, SeqBio researchers continuously develop successful tools that have an enormous biological impact. A popular approach that fits both the accelerated development timeline and sidesteps the need for TA3 is to reduce the problem to one from a toolbox of known approaches. A bioinformatician's toolbox includes black-box solvers for problems like clustering or integer linear programs; it also includes techniques like greedy algorithms or dynamic programming. This toolbox then forms the basis of incrementally designed heuristic algorithms. These are rule-based heuristics where the rules are incrementally improved by looking at where simpler heuristics fails on empirical data. This contrasts with a design driven by a mathematical understanding of the abstract problem with respect to a theoretical yardstick.

Empirical evaluation has its shortcomings but can nevertheless be very beneficial when done right. Benchmark datasets and/or community competitions have been very useful. The Genome in A Bottle consortium for example has released both a benchmark sequencing dataset and a benchmark validation dataset for the problem of structural variant detection. Similarly, competitive assessment contests, such as Assemblathon[2,6] are designed to test submitted algorithms on strategically designed datasets. Such benchmarks and competitions aim to achieve similar design and prediction goals as TA3. Empirical analysis has in fact been the major driver of algorithm design in all the three examples covered here.

Nevertheless, SeqBio suffers from serious problems that are difficult to resolve with empirical analysis, as discussed earlier. In many cases, benchmarks and competitions have not been developed, or have been developed years later than many of the algorithms (for example, in structural variation). In these cases, there is a proliferation of algorithms that perform well in the empirical validation of the authors but not in an independent evaluation. Note that the intent of the authors is not usually malicious; it is simply that the empirical validation approaches at their disposal are limited due to the lack of benchmarks. Even after the appearance of good benchmarks and competitions, algorithms that do well on these do not necessarily generalize well to other datasets; in fact, the incentives sometimes favor algorithms that overfit the data. Some assemblers, for example, are known to *perform* better on human or E. coli, which are exactly the benchmarks commonly used for design and evaluation. On the other hand, algorithms that are designed to do well on a good (that is, matching empirical observation) theoretical yardstick have the potential to be more generalizable, and the theoretical yardstick has the potential to predict algorithm performance on different datasets in a way that an empirical benchmark cannot. Thus, TA3 can overcome the problems associated with relying solely on empirical analysis.

The Challenges of Theoretical Analysis in Sequencing Bioinformatics

Based on the preceding three examples, we can speculate on what factors have made theoretical analysis more challenging for the accuracy of genome assemblers and structural variation detectors than for the memory of *k*-mer data structures. First, computer science is historically more applied to predicting memory rather than accuracy, which is more in the domain of statistics. Second, the two unsuccessful examples are closer to real data than the successful one, that is, they are more subject to the whims of poorly understood biological processes. Finally, it could simply be that compact data structures were a well-developed field before its application to SeqBio and bioinformaticians exploited that.

Moving away from the concrete examples of this article, what is it about an application domain that makes the direct application of TA3 so challenging? I can identify at least five challenges that are characteristic of SeqBio but are general enough to possibly be present in other fast-paced application domains. First, traditional worst-case analysis is in most cases too pessimistic when it comes to real data and fails to separate high-performing algorithms, which take advantage of the structure of real data, from poorly performing ones that do not. This is in fact what many empirically successful algorithms do, as they stem from a deep understanding of the data followed by heuristics to exploit its structure. Such heuristics are difficult to analyze and are unlikely to be invented when the yardstick is traditional worst-case analysis.

Second, because applied researchers require a broad inter-disciplinary skillset, they often lack the technical expertise necessary to apply more sophisticated TA3 techniques. These techniques, sometimes called beyond worst-case analysis, do in fact capture some of the complexities of real data.[44] A good example of such a technique is smoothed analysis or, more generally, semi-random models, which make the analysis more realistic by assuming there is random noise forced upon any worst-case instance. However, these advanced techniques are rarely taught as part of the core CS curricula, and applied researchers are typically exposed to theory only through introductory courses. This affects the technical complexity of the TA3 that they can perform. For example, it is a rare case that a SeqBio researcher can do a smoothed analysis of an algorithm. Being able to come up with a novel analysis technique is even more rare.

The third reason is that dataset sizes are growing at a rate faster than Moore's Law.[20] The usual justification for ignoring constants in traditional worst-case analysis is that a constant factor improvement in time or memory utilization will quickly become obsolete, as computing capacity grows exponentially. But when the size of the data is growing faster than the computing capacity, a constant factor speedup may in fact be relevant for a long time. This helps explain the success of the theoretical analysis of *k*-mer data structures, where the constant is kept.

The fourth reason is that algorithms in a fast-paced application domain are usually developed, analyzed, and applied under significant time pressure. In many cases, the algorithmic problem that a researcher is tasked with

solving is only a small part of a more complex project in the application domain (for example, biology). Because the data and its underlying technology is rapidly evolving and changing, time is of the essence. The researcher must work under time pressure to deliver a method that would analyze the data at hand and cannot afford to dedicate months of time to theoretically analyze an algorithm. There are some notable exceptions of SeqBio subareas where the data is stable enough so that the analysis and development of new algorithms has had enough time for complex TA3 techniques to emerge (for example, the edit distance problem), but these cases are rare.

The fifth reason is that there is often an incentive to publish in venues belong to the application domain rather than in CS venues. In SeqBio, for example, a paper will typically have more visibility if published in a biology journal rather than a CS bioinformatics conference. Domain scientists, however, rarely appreciate the difficulties of TA3, even if they lead to empirical breakthroughs. This especially affects early-career practitioners, who are incentivized to maximize visibility.

Because of these challenges, a useful TA3 technique must not only be predictive of empirical performance but also be easy to apply, easy to explain, and easy to understand. Thus, the theoretical analysis of algorithms in fast-paced application domains not only favors but in fact requires simplicity of the analysis technique. A simpler technique will be more trusted by domain scientists (for example, biologists), more broadly understood and applied by practitioners, more easily taught to students, and more likely included in training curricula. The challenge is to have a simple technique which nevertheless accurately captures empirical performance and is an effective yardstick for the development of empirically better algorithms.

SeqBio suffers from serious problems that are difficult to resolve with empirical analysis.

A Vision for the Future

The first step to tackling the challenges of TA3 in SeqBio is to recognize Theoretical Analysis of Applied Algorithms as its own research area, distinct from the design of the algorithms themselves. In SeqBio journals or conferences, it is currently seen as a side note of algorithm development. Even in more theoretical SeqBio venues, the value of a TA3 contribution is not always appreciated. While a breakthrough result will likely be appreciated, a paper describing incremental progress on an algorithm is much more likely to be seen favorably than an article describing substantial progress on an analysis technique. Moreover, there is generally an expectation that a SeqBio paper, even a theoretical one, delivers a novel algorithm. In most cases, this is a valid expectation, but it is not always appropriate for TA3 papers. Such publication challenges limit the formulation of TA3 subproblems and are in general detrimental to progress in the TA3 field.

The case study of SeqBio can offer insights into other fast-paced application domains. The challenges highlighted can serve as a basis for reflection in those domains and it would be interesting to compare the SeqBio experience with others. For example, the TA3 techniques needed to tackle the challenges of structural variation detection may be very different from those needed to tackle the challenges of assembly; however, by aggregating these challenges across multiple domains, we may find shared roadblocks and solutions. Recognizing TA3 as a broad research area will help with the formation of a community of like-minded researchers and all the synergies that go along with it. Unfortunately, the current fragmentation of TA3 research across multiple domains has been a bottleneck to progress.

The first step of a TA3 research program could be to survey the literature for successful applications of TA3 techniques. This article has surveyed the techniques in three sub-areas of SeqBio, but a more thorough investigation is likely to turn up some more successful applications even within SeqBio. It will be useful to also identify successful TA3 techniques from areas of bioinformatics that are not based solely on sequencing data, such as whole-genome analysis and phylogeny reconstruction. The toolbox of successful TA3 techniques can then become the starting point of further research. Moreover, it can be added to the bioinformatics curriculum in computer science.

Viewing TA3 as its own research field would allow researchers to focus on retrospective prediction of algorithm performance; that is, to develop TA3 techniques using algorithms where the empirical performance is already known. For example, when the theoretical computer science community tackled the limits of the competitive ratio analysis technique for the online paging problem, the empirically best algorithm was already known; the challenge was to find the right technique to reach the same conclusion.[44] Similarly, a distinct TA3 research program would not be afraid to tackle the question of performance prediction for short-read assembly, even though the empirically best methods have already been established.

The goal would be to develop TA3 techniques that are simple yet predictive of real-world performance. Within SeqBio, these techniques could propel it forward and enable it to respond more quickly and accurately to the rapid evolution of sequencing technology. Without investment in TA3 research, SeqBio, and other application domains will continue to be hampered by the limitations of solely empirical analysis of performance.

Acknowledgments. This article was inspired by videos for the 2014 class Beyond Worst-Case Analysis by T. Roughgarden. Thanks to R. Patro, J. Stoye, A. Tomescu, and F. Vandin for providing great feedback, M. Brudno for discussions on SVs, the twitter respondents at https://bit.ly/3I0JQRW, and J. Siren for helping me understand the role of BOSS in VG.

This material is based upon work supported by the National Science Foundation under Grants No. 2138585, 1453527, and 1931531. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM146462.

References

1. Almodaresi, F., Pandey, P., and Patro, R. Rainbowfish: A succinct colored de Bruijn graph representation. In *Proceedings in Informatics Algorithms in Bioinformatics 88*, 2017, 18:1–18:15. R. Schwartz and K. Reinert, eds. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
2. Almodaresi, F., Sarkar, H., Srivastava, A., and Patro, R. A space and time-efficient index for the compacted colored de Bruijn graph. *Bioinformatics 34*, 13 (2018), i169–i177.
3. Bankevich, A. et al. SPAdes: A new genome assembly algorithm and its applications to Single-Cell sequencing. *J. Computational Biology 19*, 5 (2012), 455–477.
4. Bowe, A., Onodera, T., Sadakane, K., and Shibuya, T. Succinct de Bruijn graphs. *WABI 7534 LNCS*. Springer, 2012, 225–235.
5. Boža, V., Brejová, B., and Vinař. T. Gaml: Genome assembly by maximum likelihood. *Algorithms for Molecular Biology 10*, 1 (2015), 1–10.
6. Bradnam, K.R. et al. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience 2*, 1 (2013).
7. Bresler, G., Bresler, M., and Tse, D. Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics 14*, S5 (2013).
8. Cairo, M. et al. The hydrostructure: A universal framework for safe and complete algorithms for genome assembly. 2020; arXiv:2011.12635.
9. Cameron, DL., Di Stefano, L., and Papenfuss, A.T. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat. Commun. 10*, 1 (July 2019), 1–11.
10. Chikhi, R., Holub, J., and Medvedev, P. Data structures to represent a set of k-long DNA sequences. *ACM Computing Surveys 54*, 1 (2021), 1–22.
11. Chikhi, R. and Rizk, G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. *WABI 7534 LNCS*. B. Raphael and J. Tang, eds. Springer, 2012,236–248.
12. Clark, S.C., Egan, B., Frazier, P.I., and Wang, Z. Ale: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics 29*, 4 (2013), 435–443.
13. Gao, S., Sung, W-K, and Nagarajan, N. Opera: Reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J. Computational Biology 18*, 11 (2011), 1681–1691.
14. Ghodsi, M. et al. De novo likelihood-based measures for comparing genome assemblies. *BMC Research Notes 6*, 1 (2013), 1–18.
15. Gilbert, J.A., Jansson, J.K., and Knight, R. The earth microbiome project: Successes and aspirations. *BMC Biology 12*, 1 (2014), 1–4.
16. Howison, M., Zapata, F., Edwards, E.J., and Dunn, C.W. Bayesian genome assembly and assessment by Markov Chain Monte Carlo sampling. *PloS One 9*, 6 (2014), e99497.
17. Hunt, M., Kikuchi, T., Sanders, M., Newbold, C., Berriman, M., and Otto, T.D. Reapr: A universal tool for genome assembly evaluation. *Genome Biology 14*, 5 (2013), 1–10.
18. Hutter, C. and Zenklusen, J. The cancer genome atlas: Creating lasting value beyond its data. *Cell 173*, 2 (2018), 283–285.
19. Kannan, S., Hui, J., Mazooji, K., Pachter, L., and Tse, D. Shannon: An information-optimal de novo RNA-Seq assembler. BioRxiv, 2016, 039230.
20. Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J., and O'Sullivan, C. The sequence read archive: A decade more of explosive growth. *Nucleic Acids Research 50*, D1 (2022), D387–D390.
21. Kingsford, C., Schatz, M., and Pop, M. Assembly complexity of prokaryotic genomes using short reads. *BMC Bioinformatics 11*, 1 (2010), 1–11.
22. Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol. 20*, 1 (June 2019), 1–18.
23. Langmead, B. and Salzberg, S.L. Fast gapped-read alignment with bowtie 2. *Nature Methods 9*, 4 (2012), 357–359.
24. Li, B. et al. Evaluation of de novo transcriptome assemblies from RNA-seq data. *Genome Biology 15*, 12 (2014), 1–21.
25. Li, D. et al. Megahit v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods 102* (2016), 3–11.
26. Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., and Sedlazeck, F.J. Structural variant calling: The long and the short of it. *Genome Biol. 20*, 1 (Nov. 2019), 1–14.
27. Marchet, C., et al. Data structures based on k-mers for querying large collections of sequencing data sets. *Genome Research 31*, 1 (2021), 1–12.
28. Mardis, E.R. DNA sequencing technologies: 2006–2016. *Nature Protocols 12*, 2 (2017), 213.
29. Marschall, T. et al. CLEVER: Cliqueenumerating variant finder. *Bioinformatics 28*, 22 (2012) 2875–2882.
30. Medvedev, P. and Brudno, M. Maximum likelihood genome assembly. *J. Computational Biology 16*, 8 (2009), 1101–1116.
31. Medvedev, P., Stanciu, M., and Brudno, M. Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods 6*, 11 (Nov. 2009), S13–20.
32. Myers, E.W. Toward simplifying and accurately formulating fragment assembly. *J. Computational Biology 2*, 2 (1995), 275–290.
33. Nagarajan, N. and Pop, M. Parametric complexity of sequence assembly: theory and applications to next generation sequencing. *J. Computational Biology 16*, 7 (2009), 897–908.
34. Narzisi, G. and Mishra, B. Comparing de novo genome assembly: The long and short of it. *PloS One 6*, 4 (2011), e19175.
35. Narzisi, G., Mishra, B., and Schatz, M.C. On algorithmic complexity of biomolecular sequence assembly problem. In *Proceedings of the 2014 Intern. Conf. Algorithms for Computational Biology*. Springer, 183–195.
36. Navarro, G. Compact Data Structures: A Practical Approach. Cambridge University Press, 2016.
37. Pandey, P., Bender, M.A., Johnson, R., and Patro, R. A general-purpose counting filter: Making every bit count. In *Proceedings of the 2017 ACM Intern. Conf. Management of Data*. ACM, New York, NY, 775–787.
38. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods 14*, 4 (2017), 417–419.
39. Rahman, A. and Medvedev, P. Representation of k-mer sets using spectrum-preserving string sets. In *Proceedings of 24th Inter. Conf. Computational Molecular Biology 12074 LNCS*. Springer, 2020, 152–168.
40. Rahman, A. and Pachter, L. CGal: Computing genome assembly likelihoods. *Genome Biology 14*, 1 (2013), 1–10.
41. Rhie, A. et al. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. bioRxiv, 2021.
42. Rhie, A. Towards complete and error-free genome assemblies of all vertebrate species. *Nature 592*, 7856 (2021), 737–746.
43. Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology 21*, 1 (2020), 1–27.
44. Roughgarden, T. Beyond worst-case analysis. *Commun. 62*, 3 (Mar. 2019), 88–96.
45. Sedgewick, R. and Flajolet, P. *An Introduction to the Analysis of Algorithms*. Pearson Education India, 2013.
46. Simpson, J.T. and Durbin, R. Efficient construction of an assembly string graph using the FM-index. *Bioinformatics 26*, 12 (2010), i367–i373.
47. Simpson, J. and Pop, M. The theory and practice of genome sequence assembly. *Annual Rev. Genomics and Human Genetics 16* (2015), 153–172.
48. Sirén, J. Indexing variation graphs. In *Proceedings of the 19th Workshop on Algorithm Engineering and Experiments*. SIAM, 2017, 13–27.
49. Titus Brown, C. Thoughts on the Assemblathon 2 paper; http://ivory.idyll.org/blog/thoughts-on-Assemblathon-2.html.
50. Tomescu, A.I. and Medvedev, P. Safe and complete contig assembly through omnitigs. *J. Computational Biology 24*, 6 (2017), 590–602.
51. Varma, A., Ranade, A., and Aluru, S. An improved maximum likelihood formulation for accurate genome assembly. In *Proceedings of the 1st IEEE Intern. Conf. Computational Advances in Bio and Medical Sciences*. IEEE, 2011, 165–170.
52. Vezzi, F., Narzisi, G., and Mishra, B. Reevaluating assembly evaluations with feature response curves: Gage and Assemblathons. *PloS One 7*, 12 (2012), e52210.
53. Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. Phenotypic impact of genomic structural variation: Insights from and for human disease. *Nature Reviews Genetics 14*, 2 (2013), 125–138.

Paul Medvedev is an associate professor in the Department of Computer Science and Engineering and the Department of Biochemistry and Molecular Biology and the Director of the Center for Computational Biology and Bioinformatics at Pennsylvania State University, University Park, PA, USA.

OPEN FOR SUBMISSIONS

ACM Journal on Computing and Sustainable Societies (JCSS)

Editor-in-Chief

Lakshminarayanan Subramanian
New York University, USA

Publishing significant and original interdisciplinary research from a broad array of fields that support the growth of sustainable societies worldwide

Inspired by the broad agenda of the United Nations Sustainable Development Goals (SDGs), the *ACM Journal on Computing and Sustainable Societies* (JCSS) aims to publish significant and original research from a broad array of computer and information sciences, social sciences, environmental sciences, and engineering fields that support the growth of sustainable societies worldwide, especially including under-represented and marginalized communities. JCSS aims to explicitly promote interdisciplinary research work including new methodologies, systems, techniques, applications, behavioral, qualitative, and quantitative studies that addresses key societal challenges including sustainability, gender equality, health, education, poverty, accessibility, conservation, climate change, energy, infrastructure and economic growth, among others. We also welcome research on the ethics of technology, especially from a critical perspective, that explores limitations and concerns with technology-led solutions for sustainable societies. The journal will be published quarterly.

For more information and to submit your manuscript, please visit acmjcss.acm.org.

Association for Computing Machinery

DOI:10.1145/3596204

To view the accompanying paper, visit doi.acm.org/10.1145/3596205 rh

Technical Perspective
Opening the Door to SSD Algorithmics

By Ramesh K. Sitaraman

SOLID-STATE DRIVES (SSDS) have revolutionized the world of non-volatile storage over the past decade. Indeed, SSDs have found their way into every nook and cranny of computing—from cell phones to powerful servers deployed in cloud datacenters. A reason for the rising popularity of SSDs is that they are faster and consume less energy than their older rivals, hard disk drives (HDDs).

Why should an algorithm designer care about the emerging dominance of SSDs? Historically, new types of computer hardware have seldom required new models of computation or new algorithmic paradigms. Much of algorithmic science uses models that are hardware independent. An enduring example is the random-access machine (RAM) that models an abstract computer that can read or write any cell in its unbounded memory in constant time. Traditionally, the RAM and related models have freed the algorithm designer from worrying about the nitty-gritty details of the underlying computer hardware.

However, hardware-independent models of computation have their limitations. As the aphorism goes, "all models are wrong, but only some are useful." Ignoring the peculiarities of the hardware can sometimes lead to models of computation that are too simplistic to be useful for algorithm design. SSDs are a case in point. SSDs store data as a set of blocks, with each block consisting of hundreds of pages. While pages can be randomly accessed for read operations, they cannot be modified in place for write operations. Further, a page must be erased before it can be rewritten and erasures only work for entire blocks, not individual pages. As a result, writing a single page in SSD can cascade into rewriting multiple pages—a phenomenon called "write amplification." For instance, when a page is modified, it must be copied over to a clean page. If a clean page is not found, a "victim" block must be selected and erased to increase the supply of clean pages—such erasure results in rewriting all valid pages in the victim block causing write amplification. Besides the performance degradation caused by write amplification, the lifetime of the SSD is also impacted since each block can only tolerate a limited number of erasures before failure. Thus, extending the lifetime of an SSD requires "wear leveling" that aims to spread the erasures evenly across blocks. Hence, a traditional model of computation that ignores the idiosyncrasies of writes and erasures in SSDs is likely too wrong to be useful.

The following paper takes a key step toward algorithm design in the SSD model of computation.

To address this critical need, the authors of the accompanying paper propose a more accurate theoretical model of SSDs that views each page as containing either valid data, invalid data (that is, stale), or no data (clean). Their model incorporates read, write, and erase operations, enabling the formal algorithmic study of key phenomena associated with SSDs, such as write amplification and wear leveling.

The paper also takes a key step toward algorithm design in the SSD model of computation. In particular, the authors derive SSD page-management algorithms for minimizing the write amplification of an arbitrary sequence of write operations. These algorithms expose the vital role "death times" play in minimizing write amplification, where the death time of a page is the next time the page is modified. Those familiar with the decades-old literature in memory cache management may recall Belady's algorithm, which achieves the maximum cache hit rate by evicting pages accessed farthest in the future. The rough analogy between the two research areas where the time of future access of pages plays a critical role in algorithm design is difficult to miss.

The formal study of algorithms and data structures for SSDs that explicitly account for write amplification and wear leveling is still in its infancy. In the current state of the art, software initially designed for other storage media is often retrofitted to work for SSDs based solely on empirical evaluation. A good model has historically been the key that opens the door to theoretical advances and algorithmic discoveries. The most significant impact of this paper is likely to come by attracting more research and researchers to the foundational study of SSD algorithmics. Given the ever-expanding role of SSDs in the modern computing universe, the benefits of such research can hardly be overstated. C

Ramesh K. Sitaraman is a Distinguished University Professor at the University of Massachusetts Amherst and Chief Consulting Scientist at Akamai Technologies, Cambridge, MA, USA.

DOI:10.1145/3596205

To view the accompanying Technical Perspective, visit doi.acm.org/10.1145/3596204

Offline and Online Algorithms for SSD Management

By Tomer Lange, Joseph (Seffi) Naor, and Gala Yadgar

Abstract

Flash-based solid-state drives (SSDs) are a key component in most computer systems, thanks to their ability to support parallel I/O at sub-millisecond latency and consistently high throughput. At the same time, due to the limitations of the flash media, they perform writes out-of-place, often incurring a high internal overhead which is referred to as write amplification. Minimizing this overhead has been the focus of numerous studies by the systems research community for more than two decades. The abundance of system-level optimizations for reducing SSD write amplification, which is typically based on experimental evaluation, stands in stark contrast to the lack of theoretical algorithmic results in this problem domain. To bridge this gap, we explore the problem of reducing write amplification from an algorithmic perspective, considering it in both offline and online settings. In the offline setting, we present a near-optimal algorithm. In the online setting, we first consider algorithms that have no prior knowledge about the input and show that in this case, the greedy algorithm is optimal. Then, we design an online algorithm that uses predictions about the input. We show that when predictions are relatively accurate, our algorithm significantly improves over the greedy algorithm. We complement our theoretical findings with an empirical evaluation of our algorithms, comparing them with the state-of-the-art scheme. The results confirm that our algorithms exhibit an improved performance for a wide range of input traces.

1. INTRODUCTION

Flash-based solid-state drives (SSDs) have gained a central role in the infrastructure of large-scale datacenters, as well as in commodity servers and personal devices. Unlike traditional hard disks, SSDs contain no moving mechanical parts, which allow them to provide much higher bandwidth and lower latencies, especially for random I/O. The main limitation of flash media is its inability to support update-in-place: after data has been written to a physical location, it has to be *erased* (cleaned) before new data can be written to it. In particular, SSDs support read and write operations in the granularity of *pages* (typically sized 4KB–16KB), while erasures are performed on entire *blocks*, often containing hundreds of pages.

To address this limitation, writes are performed *out-of-place*: whenever a page is written, its existing copy is marked as invalid, and the new data is written to a clean location. The flash translation layer (FTL), which is part of the SSD firmware, maintains the mapping of logical page addresses to physical locations (*slots*). To facilitate out-of-place writes, the physical capacity of an SSD is larger than the logical capacity exported to the application. This "spare" capacity is referred to as the device's *overprovisioning*. When the clean pages are about to be exhausted, a garbage collection (GC) process is responsible for reclaiming space: it chooses a *victim* block for erasure, rewrites any valid pages still written on this block to new locations, and then erases the block.

Block erasures typically require several milliseconds, orders of magnitude longer than page reads or writes. In addition, repeated erasures gradually wear out the flash media. As the technology advances and the capacity of flash-based SSDs increases, their *lifetime*—maximum number of erasures per block—decreases. This limitation is crucial, especially in datacenter settings, where SSDs absorb a large portion of latency-critical I/O. Thus, the efficiency of an SSD management algorithm is usually measured by its write amplification (WA)—the ratio between the total number of page writes in the system, including internal page rewrites, and the number of writes issued in the input sequence.

Mechanisms to reduce write amplification have been studied extensively by the systems community. Most mechanisms are based on *the greedy algorithm*, which chooses the block with the minimal number of valid pages (the *min-valid block*) as the victim for erasure. Several mathematical models were proposed for deriving the write amplification of the greedy algorithm as a function of overprovisioning and under various assumptions about the workload distribution.[2] The optimality of the greedy algorithm under uniform page accesses was proven in Yang et al.,[15] but it is known that its efficiency deteriorates when the workload distributions are skewed.

Garbage collection is typically conceived of as a process of selecting the optimal block to clean, whether based solely on occupancy or on predictions of future behavior. However, it seems that efficient garbage collection is also a matter of deciding where to place incoming data. Advanced techniques that separate frequently written (*hot*) pages from rarely written (*cold*) pages have been proposed. The key idea behind these techniques is that placing pages with similar write frequencies in the same block increases the likelihood that all pages within that block will be overwritten by writes with temporal proximity. Several methods for predicting page temperatures based on access context and history have proven useful in practice,[7,11] and the effectiveness of hot/cold data separation has further motivated the

The original version of this paper was published in *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2021.

development and proliferation of advanced storage hardware and interfaces.

1.1. Our model

We assume that P logical pages are stored in S physical SSD slots, where pages and slots are of the same size. The slots are organized in B blocks, each block consists of n slots. The relation between the physical capacity and the logical capacity is expressed through a fixed *spare factor*, which is defined by $\alpha = \frac{S-P}{S}$. Consider, for example, the SSD in Figure 1. In this toy example, the SSD consists of five blocks, four slots each, and stores ten logical pages. Thus, the spare factor is $\alpha = \frac{20-10}{20} = \frac{1}{2}$.

The most recent version of a page is considered *valid*, whereas all other versions are considered *invalid*. At every time t, each slot in the SSD must be in one of the following deterministic *states*: *valid*, *invalid,* or *clean*. A slot is called *valid* if it stores a valid version of a page. A valid slot that stores page p becomes *invalid* when an update request for p arrives. Following a clean operation of block b, each slot in b is rendered *clean*.

At every time t, a new update request for some page p arrives, and the new page must be *written* to a clean slot. When cleaning block b, each valid page stored in b must be *rewritten* to a clean slot. Thus, the SSD manager should make two types of decisions: (1) which block to clean and when, and (2) which slot to allocate to each page. In Figure 1, an update request for page p_1 arrives, and the SSD manager writes it to a slot in b_3 (note that p_1 could also be written to a slot in b_2 or b_4). Then, the SSD manager decides to clean b_5, and p_2, p_3 are rewritten to b_4, b_3, respectively.

Problem statement. In the *SSD Management Problem*, we are given an input in the form of a *request sequence* $\sigma = (\sigma_1, \sigma_2, ..., \sigma_T)$ of length T. The objective is to minimize the number of rewrites during the service of σ. We evaluate the performance of an algorithm $\mathcal{ALG}$ given an input σ using the *write amplification* metric, which is defined by

$$WA_{\mathcal{ALG}}(\sigma) = \frac{T+Z}{T} = 1 + \frac{Z}{T}$$

where Z is the number of rewrites performed by $\mathcal{ALG}$ during the service of σ.

1.2. Our contribution

When developing online algorithms for SSD management, a fundamental issue is what kind of "extra" information about the input sequence is useful for improving performance. A relatively recent study[6] observes that pages should ideally be grouped in blocks according to their *death times*, that is, pages that are about to be accessed at nearby times in the future should be placed in the same block. A more recent study applies this principle in a heuristic FTL.[3] We call this rule *grouping by death time*. When all the pages in a block become invalid at approximately the same time, write amplification can be minimized by choosing the block whose pages have all been invalidated as the victim block. In a sense, this is precisely what hot/cold data separation strives to achieve; however, hot/cold separation can be viewed as grouping by *lifetime*, while two pieces of data can have the same lifetime (i.e., hotness) but distant death times. While hot/cold data separation has become standard practice, grouping by death time is only now coming to the attention of the systems research community.

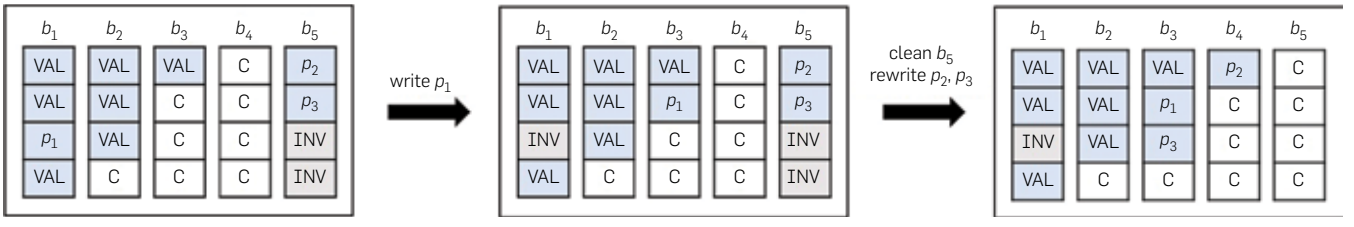
Death times can be predicted using machine-learning techniques[3]; however, despite the success of machine-learned systems across many application domains, they tend to exhibit errors when deployed. A natural question is then how equipping online algorithms with a machine-learned oracle can benefit SSD performance. In particular, we look for an algorithm that: (1) performs better as the oracle's error decreases; and (2) behaves at least as good as the best online algorithm that has no knowledge of the future, independently of the oracle's performance on a particular instance.

The above is very much in line with recent trends of applying machine learning predictions to improve the performance of online algorithms.[8, 9, 12] The work of Lykouris and Vassilvitskii[8] is particularly relevant to our setting. They studied the classic paging problem assuming the availability of a machine-learned oracle that for each requested page, predicts the next time it will appear in the request sequence. They design a competitive online algorithm using these predictions and analyze its competitive factor as a function of the oracle's error. When predictions are pretty accurate, their algorithm circumvents known lower bounds on the competitive factor of online paging.

In this work we study the SSD management problem from an algorithmic perspective, considering it in both offline and online settings. A key parameter used in our bounds is the *spare factor*, denoted by α. This factor expresses the relation between the physical capacity of the SSD and the logical capacity. We now present our main technical results.

Myopic algorithms. In Section 2, we study online algorithms having no prior knowledge about the request sequence, for example, the well-known greedy algorithm.

Figure 1. Example SSD with five blocks and 10 logical pages. First, page p_1 is written to a clean slot, and its previous location is invalidated. Then, b_5 is cleaned, and pages p_2, p_3 are rewritten.



We analyze this algorithm and prove that no deterministic algorithm can provide better performance in this setting.

Offline setting. In Section 4, we consider an offline setting in which the request sequence is known in advance. We propose a novel placement technique that uses information about the death times of the pages to improve the performance of the SSD. We apply this technique and design a near-optimal algorithm whose write amplification given any input is not greater than $1 + \varepsilon$. This result highlights the fact that death times are very useful for minimizing write amplification. Moreover, this is the first time that an offline algorithm for SSD management is analyzed, and from a systems perspective, having a near-optimal offline algorithm is very useful for benchmarking (similarly to the role that Belady's algorithm plays in paging).

We also consider the hardness of the offline SSD management problem. While it is not known whether this problem is solvable in polynomial time, we take a step in resolving this issue and link it to the problem of *fragmented coloring* on interval graphs, which has been previously studied in Diwan et al.[5] Even though the complexity of the latter problem remains open, we extend previous results by providing an efficient algorithm for the special case in which the number of colors is fixed. Details can be found in the full version of the paper.

Online setting. In Section 5, we consider a middle ground setting in which we assume the availability of an oracle that predicts the death time of each accessed page. We design an online algorithm whose performance is given as a function of the prediction error η. Specifically, we prove that for any input σ, our algorithm satisfies:

$$WA_{online}(\sigma, \eta) \leq \min\left(1 + O\left(\frac{\eta}{T}\right), \frac{1}{\alpha}\right) + \varepsilon.$$

Note that our algorithm exhibits a graceful degradation in performance as a function of the average prediction error. Furthermore, our algorithm is *robust*, that is, its performance is never worse than that guaranteed by the greedy algorithm, even when the prediction error is large. When predictions are perfect ($\eta = 0$), the write amplification is bounded by $1 + \varepsilon$, similar to the offline setting. However, ε here has a larger value, since the online model is less informative.

Experimental evaluation. In Section 7, we empirically compare our online and offline algorithms to state-of-the-art practical SSD management with hot/cold data separation. Our results confirm that our algorithms exhibit an improved performance for a wide range of input traces; the highest benefit is achieved for traces with modest or no skew, which are the worst-case inputs for algorithms with hot/cold data separation.

Summary. The abundance of system-level optimizations for reducing write amplification, which is usually based on experimental evaluation, stands in stark contrast to the lack of theoretical algorithmic results in this problem domain. We present the first theoretical analysis and systematic evaluation of algorithms that observe the rule of grouping by death time, laying the ground for the development of a new class of algorithms that the systems community has not yet considered. Our results may also motivate the theory community to address the algorithmic aspects of SSD-related design challenges.

2. MYOPIC ALGORITHMS

We consider here *myopic* algorithms, which are online algorithms that have no access to information about future requests. Intuitively, in this setting, cleaning a block that minimizes the number of valid slots is advantageous since it requires fewer rewrites and frees more slots for reuse. In this section, we prove that this intuition is indeed correct.

The well-known greedy algorithm for SSD management works as follows: it waits until the SSD becomes full, and then cleans the *min-valid* block, that is, the block that minimizes the number of valid slots. We now provide a simple upper bound on the write amplification guaranteed by the greedy algorithm on any input and claim that no deterministic myopic algorithm can provide a better guarantee. The proofs for the following theorems can be found in the full version of this paper.

THEOREM 1. *For any input* σ, $WA_{greedy}(\sigma) \leq \frac{1}{\alpha}$.

THEOREM 2. *Let* $\mathcal{ALG}$ *be a deterministic myopic algorithm for SSD management. Then, there exists an input* σ *such that* $WA_{\mathcal{ALG}}(\sigma) \geq \frac{1}{\alpha}$.

2.1. Randomized lower bound

Randomization is a common approach to circumvent an adversary that "deliberately" feeds a bad input to a deterministic algorithm. A natural question is whether randomization can be used to circumvent the lower bound presented in Theorem 2 against an adversary that is oblivious to the random choices made by the algorithm. Surprisingly enough, it turns out that the power of randomization in the context of myopic algorithms for SSD management is very limited, as follows.

Yao's principle[16] states that the expected cost of a randomized algorithm on a worst-case input is lower bounded by the expected performance of the best deterministic algorithm for a given probability distribution over the input. This principle is a powerful tool for proving lower bounds on the performance of randomized algorithms.

In previous work, Desnoyers[4] investigated the expected write amplification of several algorithms under the assumption that accesses are uniformly distributed between the pages. In particular, he shows that the write amplification of the greedy algorithm under such a distribution is approximately $1/2\alpha$. This result is consistent with prior work that evaluated the performance of the greedy algorithm empirically.[1] The optimality of the greedy algorithm under a uniform distribution assumption was proven in Yang et al.[15] Hence, we have by Yao's principle that no randomized myopic algorithm can guarantee an expected write amplification that is less than $1/2\alpha$. This leaves only a small gap for improvement in the randomized setting.

3. OUR APPROACH

The inherent limitation of myopic algorithms in general, and specifically the greedy algorithm, is due to two main reasons. The first is that choosing a victim block according to the number of valid slots may not be an optimal heuristic. For example, it is reasonable to delay the cleaning of a block whose pages are about to be accessed in the near future,

even if it contains a small number of valid slots. The second reason, on which we focus in this work, is that myopic algorithms cannot make informed decisions on where to write pages. From the point of view of a myopic algorithm, all pages are semantically identical. Obviously, this limitation can be exploited by an appropriate adversary. In this section, we propose the notion of *death time* as a measure that can distinguish between pages. Then, we demonstrate how grouping pages with similar death times in the same block can benefit the performance of the SSD. The ideas presented in this section form the basis for both offline and online algorithms presented in the sequel.

Definition 1. The death time of page p at time t, denoted by $y(p, t)$, is the time of the next access to p after time t.

When a block stores pages with different death times, there is a time window between the first and the last page invalidations within which the block contains both valid and invalid slots. We call such a time window a *zombie window* and a block in a zombie window a *zombie block*.[6] In general, larger zombie windows lead to increased odds of a zombie block being chosen as a victim, increasing the number of rewrites.

Zombie windows can be reduced if pages with similar death times are placed in the same block. This rule is referred to as *grouping by death time*.[6] However, algorithms relying on the death times of pages must be equipped with information about the future. This information can be exact (Section 4) or estimated by a prediction oracle (Section 5). We now introduce the basic principle of our death-time-aware placement technique.

Definition 2. Let $\mathcal{R}$ be a set of (*page, time*) tuples, let $\mathcal{B}$ be a set of blocks, and let f be a function that given a tuple $(p, t) \in \mathcal{R}$, outputs the predicted death time of page p at time t. Consider a function A that maps each tuple $(p, t) \in \mathcal{R}$ to a block in $\mathcal{B}$. We say that A allocates space *according to* f if for any $(p_1, t_1), (p_2, t_2), (p_3, t_3) \in \mathcal{R}$ such that $f(p_1, t_1) < f(p_2, t_2) < f(p_3, t_3)$, if $A(p_1, t_1) = A(p_3, t_3) = b$ then $A(p_2, t_2) = b$ as well.

Figure 2. An allocation according to the exact death times. Each entry in the third row contains the death time (DT) of the corresponding page. Since $(p_1, 1)$ and $(p_4, 5)$ die first, they are mapped to the same block.

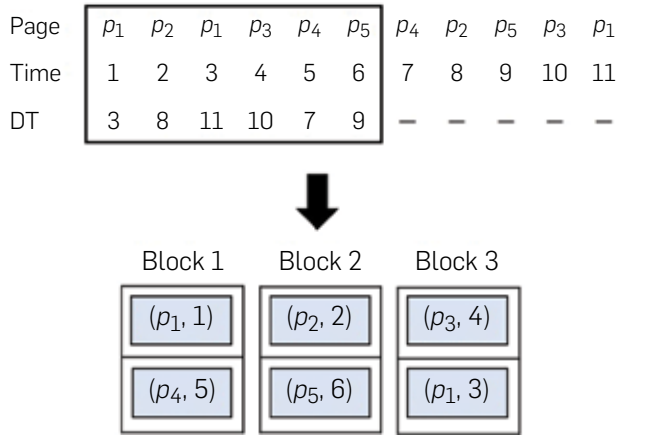


Figure 2 demonstrates an allocation according to the (exact) death times. The first two rows describe the request sequence. Each entry in the third row contains the death time (DT) of the corresponding page. For example, p_1 is written at time $t = 1$, and this version of p_1 dies at time $t = 3$ when an update request for p_1 arrives. In this simple scenario, the set $\mathcal{B}$ consists of three blocks, two slots each. The slots are allocated to the first six accessed pages according to their death times. For example, since $(p_1, 1)$ and $(p_4, 5)$ die first, they are mapped to the same block.

We finish this section by introducing the idea lying at the core of our algorithms. A set of blocks whose slots are allocated to pages according to their (predicted) death times is called a *batch*, and the total number of batches is denoted by Δ. Observe that when allocating slots according to the *exact* death times of the pages, each batch may contain at most one zombie block. Hence, Δ can function as a measure of the disorder induced by the page locations: we expect that the greater the value of Δ, the greater the number of zombie blocks. The idea is that we maintain low values of Δ by reordering all the pages when their value becomes too high. Since such a reordering requires many rewrites, we also ensure that this operation is done infrequently.

4. THE OFFLINE SETTING

In this section, we address the SSD management problem in an offline setting in which the request sequence is fully known in advance. We design an algorithm that applies the placement technique introduced in Section 3. Then we prove that our algorithm performs a negligible number of rewrites (relative to the input length), highlighting both the importance of the placement policy (rather than the victim selection strategy) and the usefulness of death times for minimizing write amplification.

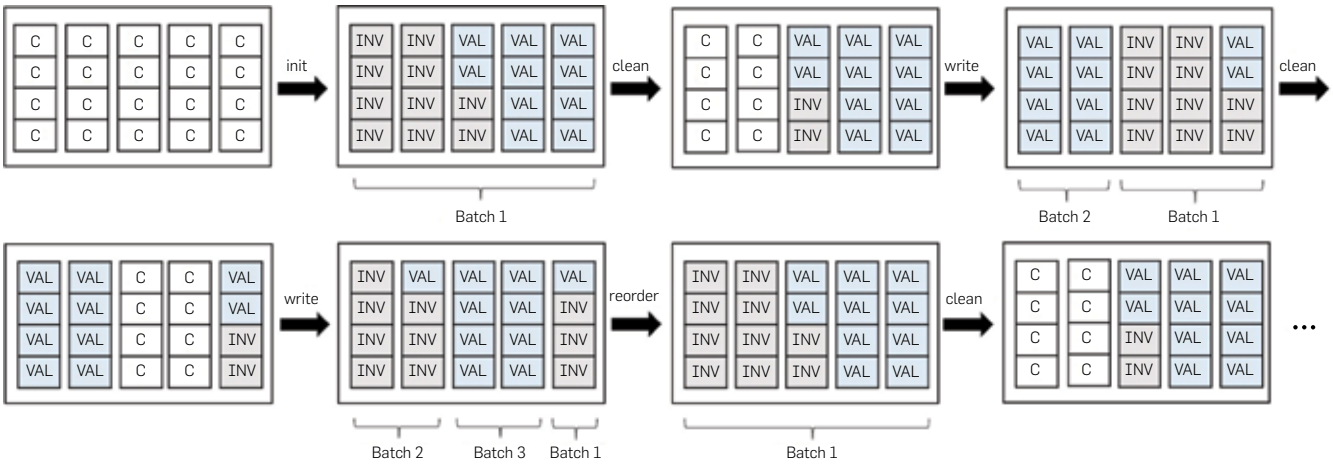
4.1. The offline algorithm

We now describe the proposed offline algorithm. We assume that the SSD is initially clean. In the initialization step, we fill the SSD with the first S accessed pages according to their death times. Now, we perform a sequence of *iterations*. In each iteration, we first choose the set $\mathcal{B}_{victim}$ of *completely* invalid blocks as victims (that is, blocks containing only invalid slots), and clean them. Then, the subsequent accessed pages are written to $\mathcal{B}_{victim}$ according to their death times. As soon as the device becomes full again, a new iteration begins.

Observe that in each iteration, a new allocation is calculated. Thus, after j iterations, the number of batches can be at most j (note that the number of batches can be smaller than the number of iterations if the pages of an entire batch have been invalidated in subsequent iterations). In order to keep Δ low, when it exceeds the predefined value $\alpha \cdot B$, a *reordering step* is invoked. In this step, all pages are reordered in $\frac{P}{n}$ blocks according to their death times and Δ is reset to 1. The iterations performed between two reordering steps form a *phase*. From now on, iteration j in phase i is simply referred to as iteration (i, j).

Figure 3 presents a schematic example of the first phase in the offline algorithm. All slots are initially clean. In the initialization step, the pages are written to the SSD according to their death times, such that pages that die first are written to the leftmost block. Note that due to overprovisioning, some slots in the left blocks are invalidated already during the initialization.

Figure 3. Schematic illustration of the first phase in the offline algorithm. Each iteration consists of a clean step and a write step. As soon as $\Delta \geq 2.5$, a reordering step is invoked.



When the device becomes full, the first phase begins. In each iteration, all completely invalid blocks are cleaned. Then, the next accessed pages are written to the clean slots, forming a new batch. As soon as $\Delta \geq \alpha B$, a reordering step is invoked.

Note that in general, reordering the pages according to their death times might require a huge number of rewrites. However, under specific circumstances, the reordering process can be done efficiently. In the full version of this paper, we propose a simple procedure that rewrites each page only once. A formal version of the algorithm is given in Algorithm 1.

Algorithm 1: offline

```
   input: A request sequence σ
 1 initialization step;
 2 for each phase i = 1, 2, ... do
 3     for each iteration j = 1, 2, ... do
 4         if Δ ≥ αB then break;
 5         choose the set B_victim of blocks containing
             only invalid slots as victims;
 6         clean the victim blocks;
 7         fill B_victim with the next accessed pages,
             written according to their death times;
 8     end
 9     reordering step;
10 end
```

THEOREM 3. *For any input σ,*

$$WA_{offline}(\sigma) \leq 1 + O(1/B).$$

PROOF. We assume that in each reordering step each page is rewritten at most once, with a total of up to P rewrite operations. Since rewrites occur only in the reordering step, our analysis focuses on examining the frequency with which this step occurs.

Recall that when the device is full, the number of invalid slots in the whole device is exactly $\alpha \cdot S$. By allocating slots according to the exact death times of the pages, we ensure that each batch can contain at most one zombie block. Hence, the number of completely invalid blocks when the device is full is at least $\alpha B - \Delta$. This way, our condition for ending a phase guarantees that for any phase i, at least $(\alpha B - j) \cdot n$ new requests are served during iteration (i, j). Thus, the total number T_{ph} of requests served in each phase satisfies:

$$T_{ph} \geq \sum_{j=1}^{\alpha B} (\alpha B - j) n = \frac{1}{2} n (\alpha B)^2.$$

We obtain that for any input σ,

$$WA_{offline}(\sigma) \leq 1 + \frac{P}{T_{ph}} = 1 + O(1/B),$$

which completes the proof. □

5. ONLINE SETTING WITH PREDICTIONS

In this section, we extend our discussion to online algorithms that have access to death-time predictions and present an algorithm based on our placement technique with improved performance.

5.1. APPROACH

We first discuss several challenges arising when trying to apply our placement technique in the online setting and present our approach for addressing them.

Challenge 1: In the online setting, information about the death times of the pages is not available.

Approach: We assume the availability of a prediction oracle, such that each request arrives with its predicted death time. Such an oracle can be derived from machine learning prediction techniques; however, since it might exhibit errors when deployed, the performance of the proposed algorithm is given as a function of the oracle's error. We emphasize that our algorithm is completely agnostic to the way this oracle works and treats it as a complete black box.

Challenge 2: Since predictions arrive together with requests, we do not have the predictions for the next accessed pages in advance. Therefore, an allocation based on their death times cannot be calculated.

Challenge 3: When prediction errors occur and zombie blocks must be chosen as victims, the pages they store should be reordered according to their predicted death times. However,

reordering these pages might require an excessively large number of rewrites, harming the performance of the algorithm.

Approach: In order to address Challenges 2–3, we augment our model, assuming the availability of an auxiliary memory of size M. We assume that this memory is much faster than the SSD, allowing us to neglect implications concerning its management. In practice, such memory is typically attached to the SSD controller and is used for the FTL as well as for write buffering, although its size is limited.

We (logically) partition the memory into two *zones*: a *buffer zone* of size αM and a *victim zone* of size $(1-\alpha)M$. When a zombie block is cleaned, the pages it stores are retrieved to the *victim zone*. When a new request for page p arrives, instead of writing p to the SSD immediately, it is first stored in the *buffer zone*. As soon as the buffer zone becomes full, an allocation is calculated, and then all pages stored in the memory are moved to the SSD according to their predicted death times.

5.2. The online algorithm

Our goal is to design an algorithm whose performance improves as the prediction error decreases. We achieve this objective by treating the oracle's output as the truth; this corresponds to a policy that allocates slots to pages according to their predicted death times. At the same time, we are interested in a *robust* algorithm, whose performance guarantees are comparable to those of the best online algorithm (i.e., the greedy algorithm), regardless of the performance of the oracle. In order to limit the effect of prediction errors, the predictions are used only in the placement policy. Specifically, we apply a greedy selection of min-valid blocks as victims, regardless of the predictions.

We now provide a short description of the proposed algorithm, whose formal description appears in Algorithm 2. Similarly to the offline case, our algorithm operates in phases. Each phase contains several iterations and terminates with a reordering of the pages on the entire SSD. Each iteration consists of: (1) a *service process*, in which incoming requests are served; (2) a *garbage collection process*, in which victim blocks are chosen and cleaned; and (3) a *placement process*, in which pages are moved from the memory to the SSD.

Algorithm 2: online

input: A request sequence σ and a confidence parameter ρ

1 **for** *each phase* $i = 1, 2, \ldots$ **do**
2 **for** *each iteration* $j = 1, 2, \ldots$ **do**
3 **if** $\Delta \geq \rho \cdot \alpha B$ **then** break;
4 serve the next requests by storing new pages in the *buffer zone*, until it becomes full;
5 choose the set $\mathcal{B}_{victim}$ of the B_M min-valid blocks as victims (ties broken arbitrarily);
6 clean the victim blocks and retrieve the pages they store to the *victim zone*;
7 move the pages from the memory to $\mathcal{B}_{victim}$ according to their predicted death times;
8 **end**
9 reordering step;
10 **end**

In the *service process* (line 4 in Algorithm 2), when an update request for page p arrives, we store the new version of p in the *buffer zone*. If the previous version of p was stored in the memory, we overwrite it; otherwise, we invalidate the corresponding slot in the SSD. This process continues until the *buffer zone* becomes full, and then garbage collection begins. In the *garbage collection process* (lines 5–6 in Algorithm 2), we first choose the set $\mathcal{B}_{victim}$ of $B_M = \frac{M}{n}$ min-valid blocks as victims (note that the physical capacity of the victim blocks is equal to that of the entire memory). We then clean these blocks, and retrieve valid pages they store to the *victim zone* (note that, since we clean min-valid blocks, the *victim zone* is guaranteed to have enough room for these pages). In the *placement process* (line 7 in Algorithm 2), we move the pages from the memory to the SSD according to their predicted death times.

When the condition for ending a phase is satisfied, we perform a reordering step. An efficient implementation for this step can be found in the full version of this paper. In contrast to the offline case, the online algorithm gets a fixed *confidence parameter* $0 < \rho < 1$ as input. This parameter should express the user's degree of confidence in the oracle's predictions, where higher confidence is expressed through higher values of ρ. The condition for ending a phase depends on the value of ρ, where higher values of ρ lead to longer phases. As a consequence, higher values of ρ increase the dependency of the performance on the magnitude of prediction errors and decrease the overhead imposed by the reordering steps. In the extreme case, if ρ is very close to 1, then we no longer bound the performance of the algorithm as a function of the error's magnitude. On the other hand, if ρ is very close to 0, then the algorithm performs an infinite sequence of reordering steps without serving any request, resulting in an unbounded write amplification.

5.3. Performance modeling

The above adaptations require us to remodel the performance of algorithms. Obviously, write amplification guarantees may now be given as a function of the oracle's error and the memory size.

Error model. Denote by $\hat{y}(p, t)$ the predicted death time of page p at time t. We define the prediction error on a specific request (p, t) by $\eta_t = |y(p, t) - \hat{y}(p, t)|$, and the total error by $\eta = \sum_t \eta_t$. Note that we do not make any assumptions about the oracle's error; specifically, our results will be parameterized as a function of the error magnitude, and not its distribution. We note that in practice, our results will be with respect to the *observed* value of η (i.e., the prediction errors that "interfere" with the SSD management), rather than its actual value, which might be much higher.

Objective. Observe that when page overwrites occur, some page versions are never written to the SSD. Hence, we redefine the write amplification of an algorithm $\mathcal{ALG}$ given input σ by:

$$WA_{\mathcal{ALG}}(\sigma) = \frac{\#system\_writes}{\#user\_writes}$$

where the *system writes* counts every write of a page to the SSD and $\#user\,writes = T$. In particular, note that now write amplification can be smaller than one (due to page overwrites in

memory). In any case, under these conventions, minimizing the write amplification remains our objective.

We obtain the following bound on the write amplification of the proposed algorithm.

THEOREM 4. *For any input* σ,

$$WA_{online}(\sigma, \eta) \leq \min\left(1 + O\left(\frac{\eta}{ST}\right), \frac{1}{\alpha}\right) + O(1/B_M).$$

The proof of this theorem is in the spirit of Theorem 3; however, it involves more details due to the dependence on prediction errors. A complete proof can be found in the full version of this paper.

6. EXPERIMENTAL SETUP

We complement our theoretical findings with an experimental evaluation. Our goal is to demonstrate how our algorithms compare with the state-of-the-art practical algorithms for SSD management. For the purpose of this evaluation, we built a simple SSD simulator similar to the one used in Yadgar et al.[14]

6.1. Algorithms

We implemented the three algorithms, greedy, offline, and online, according to their description in Sections 2, 4, and 5, respectively. We ran our online algorithm with three values of ρ: 0.25, 0.5, and 0.75. For brevity, here we discuss only the results for $\rho = 0.5$.

We also implemented an SSD management algorithm with hot/cold data separation (*hotcold*). Hotcold classifies pages as hot, warm, or cold, based on the death-time predictions used by our online algorithm. It writes the pages into disjoint sets (partitions) of blocks, with Greedy SSD management applied within each partition. The sizes of the partitions are adjusted dynamically to the input, by allocating blocks to partitions on demand. The dynamic allocation of blocks to partitions results in dynamic partition sizes, which is suitable when using death times: a page can be classified with a different temperature when it is written in different requests. This is on par with how practical hot/cold data separation schemes classify page temperatures dynamically.

6.2. Memory usage

We extended the standard greedy and hotcold algorithms to use the available memory as a write-back cache. The greedy algorithm, which does not rely on death time predictions, manages the cache with least-recently-used (LRU) replacement. Namely, in case of cache "hit"—a copy of the page is already in the memory, and it is overwritten. In case of cache "miss," the least-recently accessed page in the memory is written to the SSD and evicted from the memory. The valid copy of the requested page on the SSD is invalidated, and its new copy is written in the memory. Hotcold uses the memory in a similar way, but gives priority to hot and warm pages in the memory: when the memory is full, a new page can cause an old page to be evicted only if its own temperature is equal to or warmer than that of the old page. For a fair comparison, we also implemented an extension of the offline algorithm that uses memory as a write buffer, using Belady's law for page eviction.

6.3. Predictions and errors

We pre-processed all traces in order to annotate each page request with its death time. To evaluate the efficiency of our algorithm, we also created erroneous versions of those annotations. Specifically, we annotated each request σ_t for page p with a random value that is uniformly distributed between t and its correct death time $y(p, t)$. Intuitively, we expect that in practice, prediction errors will be larger for pages whose next access is farther in the future.

6.4. Input traces

We used two trace types: synthetic and real. The synthetic traces were generated according to uniform and Zipf distributions. We also used a synthetic trace which sequentially writes entire files. The real traces were taken from servers at MSR Cambridge,[10] available via SNIA.[13] These traces are extensively used by the storage systems community to evaluate caching, scheduling, and SSD management algorithms. The real-world traces exhibit a high skew, that is, a small percentage of their pages is responsible for a large portion of their accesses.

6.5. SSD parameters

In our experiments, we compared the performance of the above algorithms on an SSD with 128-page blocks and varying spare factors and memory sizes. For brevity, we present here only the results for $\alpha = 0.1$. The full details of our implementation and trace characteristics, as well as additional experimental results, can be found in the full version of this paper.

7. EXPERIMENTAL RESULTS

7.1. Accurate predictions

In our first experiment, we compared our proposed algorithms to the two practical online algorithms greedy and hotcold, on all traces, with increasing memory sizes and accurate death time predictions. Figure 4 shows the write amplification of these algorithms. The results confirm the common knowledge that hot/cold separation can significantly reduce write amplification with respect to greedy, and this reduction increases with the skew in the workload. Our online algorithm outperforms greedy on all the input traces and all settings, and it usually outperforms hotcold as well. The offline algorithm consistently achieves the lowest write amplification.

The write amplification is highest for the Uniform trace: since all the pages have the same update frequency, the probability of a cache hit is low for all algorithms. Our online algorithm exhibits a significant improvement over both greedy and hotcold for this trace, and its write amplification is higher than that of the offline algorithm by only 7%–23%.

The skewed accesses in the Zipf and real traces (prxy_0, proj_0, and src1_2) imply that the memory can absorb a high portion of the requests, even if it only stores the most recently used pages. Indeed, the write amplification is lower than 1 in most parameter combinations for those traces. Our online algorithm is mostly better than greedy and hotcold; its advantage is larger in settings with modest memory sizes, where the effectiveness of the memory as a cache is limited.

The Sequential trace presents our online algorithm with a challenge. The large sequential writes always generate at least

one block containing only invalid slots; hence, the greedy and hotcold algorithms simply erase one such block at a time, which is sufficient for serving the entire trace without rewrites. The online algorithm, on the other hand, always chooses B_M blocks as victims, which sometimes results in unnecessary rewrites (especially for the larger memory sizes). For this reason, the three online algorithms are comparable on this trace.

7.2. The effect of prediction errors

In the next experiment, we evaluate the performance of our online algorithm using erroneous death-time predictions. We omit greedy from this experiment, because it does not use the death time annotations in the input, and is therefore not affected by prediction errors. Figure 5 shows the write amplification of hotcold and our online algorithm. As we expected, both algorithms are sensitive to prediction errors, because they rely on the predictions in order to (1) organize the pages in the SSD; and (2) select pages for eviction. Our online algorithm is more sensitive to errors than hotcold because its allocation decisions are made with finer granularity.

Quantitatively, the prediction errors roughly double the write amplification of our online algorithm on both Uniform and Zipf traces. Prediction errors also affect our online algorithm with the Sequential trace: they cause it to spread pages from the same file on more blocks than necessary. We note, however, that sequential accesses in practice will not experience this extreme error scenario, as this is a relatively easy access pattern to predict. Hotcold is hardly affected by prediction errors in this trace because, in the worst case, a file will be split into at most three continuous slot ranges, one in each temperature.

On the prxy_0 trace, our online algorithm outperforms hotcold even with prediction errors. The other real traces, proj_0 and src1_2, contain a significant percentage of sequential accesses; indeed, real-world workloads often include a small set of hot pages and long sequential updates of large files or logs. For sequential access, the effectiveness of the memory as a cache is limited, which explains the higher write amplification of all the algorithms for those traces. The write amplification of our online algorithm is mostly higher than that of hotcold, which performs better on sequential accesses in the presence of prediction errors.

7.3. Consequences

Our offline algorithm achieves the lowest write amplification in all simulations, indicating the effectiveness of our death-time-aware placement technique. Its highest benefit is achieved for traces with modest or no skew, which are the worst-case inputs for the greedy and hotcold algorithms. Our online algorithm emphasizes worst-case guarantees, resulting in an excessive cost in the case of sequential accesses, which are handled naturally by the log-structured organization of greedy. On the other hand, in traces that do not present

Figure 4. The write amplification of greedy, hotcold, and our online and offline algorithms with varying memory sizes and accurate death time predictions. The memory size is given as the percentage of the SSD's physical size.

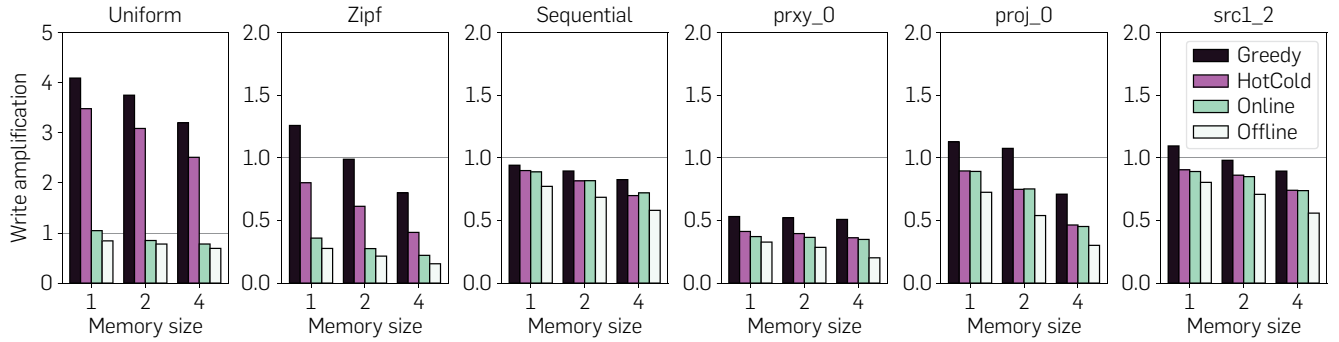
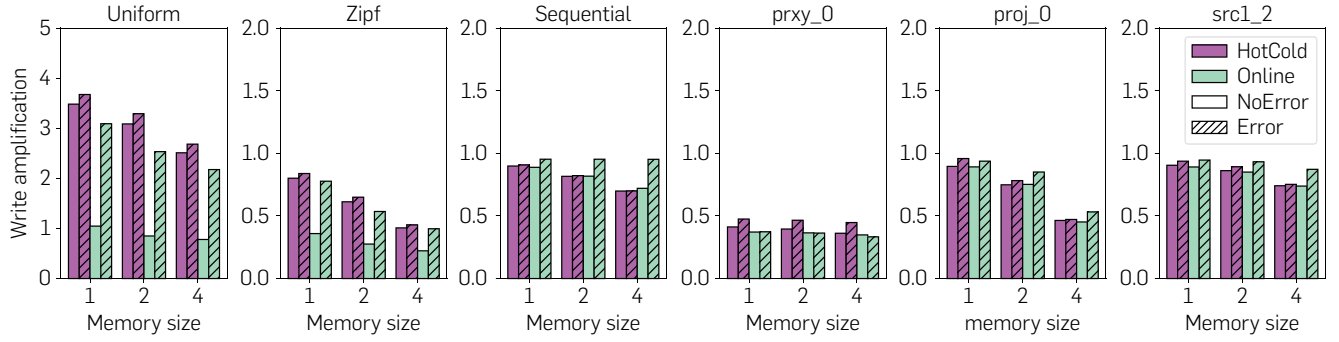


Figure 5. The write amplification of hotcold and our online algorithm with varying memory sizes with (Error) and without (NoError) prediction errors. The memory size is given as the percentage of the SSD's physical size.



frequent sequential patterns, our online algorithm mostly outperforms other algorithms, especially when predictions are accurate — emphasizing that minimizing the oracle's error is important in order to take advantage of our technique.

8. CONCLUSION AND OPEN PROBLEMS

In this paper, we initiated a comprehensive study of the SSD management problem. We first defined the problem formally and then explored it in both offline and online settings from a worst-case algorithmic perspective. Our paper opens up several intriguing research directions. The first one is resolving the complexity of the offline SSD management problem. We suspect that this problem is NP-hard, and thus finding an optimal approximation for it is an interesting avenue for further research. Another interesting direction is improving the dependency of our online algorithm on both the prediction error and the size of the auxiliary memory. Empirically, our online algorithm successfully deals with a wide range of input traces, and hence can be integrated with state-of-the-art algorithms to improve SSD performance. C

References

1. Agarwal, R., Marrow, M. A closed-form expression for write amplification in NAND flash. In *2010 IEEE Globecom Workshops* (2010), IEEE, Miami, FL, USA, 1846–1850.
2. Bux, W., Iliadis, I. Performance of greedy garbage collection in flash-based solid-state drives. *Perform. Eval. 67*, 11 (Nov. 2010), 1172–1186.
3. Chakraborttii, C., Litz, H. Reducing write amplification in flash by death-time prediction of logical block addresses. In *Proceedings of the 14th ACM International Conference on Systems and Storage* (2021), ACM, Haifa, Israel.
4. Desnoyers, P. Analytic models of SSD write performance. *ACM Trans. Storage 10*, 2 (Mar. 2014) 1–25.
5. Diwan, A., Pal, S., Ranade, A. Fragmented coloring of proper interval and split graphs. *Discrete Appl. Math. 193* (2015), 110–118.
6. He, J., Kannan, S., Arpaci-Dusseau, A.C., Arpaci-Dusseau, R.H. The unwritten contract of solid state drives. In *Proceedings of the 12th European Conference on Computer Systems (EuroSys '17)* (2017), ACM, NY, 127–144.
7. Hsieh, J.-W., Kuo, T.-W., Chang, L.-P. Efficient identification of hot data for flash memory storage systems. *ACM Trans. Storage 2*, 1 (Feb. 2006), 22–40.
8. Lykouris, T., Vassilvitskii, S. Competitive caching with machine learned advice. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of Proceedings of Machine Learning Research* (2018), PMLR, Stockholm, Sweden, 3302–3311.
9. Mitzenmacher, M., Vassilvitskii, S. Algorithms with predictions. In *Beyond the Worst-Case Analysis of Algorithms*. T. Roughgarden, ed. Cambridge University Press, Cambridge, UK, 2020, 646–662.
10. Narayanan, D., Donnelly, A., Rowstron, A. Write off-loading: Practical power management for enterprise storage. *ACM Trans. Storage 4*, 3 (Nov. 2008), 1–23.
11. Park, D., Du, D.H. Hot data identification for flash-based storage systems using multiple Bloom filters. In *27th IEEE Symposium on Mass Storage Systems and Technologies (MSST)* (2011) IEEE, Denver, CO, USA.
12. Purohit, M., Svitkina, Z., Kumar, R. Improving online algorithms via ML predictions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018* (December 3–8, 2018, Montrèal, Canada) (2018), Curran Associates, Inc., Montreal, Canada, 9684–9693.
13. SNIA IOTTA Trace Repository. MSR Cambridge Traces, 2007. http://iotta.snia.org/traces/block-io/388.
14. Yadgar, G., Gabel, M., Jaffer, S., Schroeder, B. SSD-based workload characteristics and their performance implications. *ACM Trans. Storage 17*, 1 (Jan. 2021) 1–26.
15. Yang, Y., Misra, V., Rubenstein, D. On the optimality of greedy garbage collection for SSDs. *SIGMETRICS Perform. Eval. Rev. 43*, 2 (Sept. 2015), 63–65.
16. Yao, A.C.-C. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (SFCS)* (1977), IEEE, Los Alamitos, CA, USA, 222–227.

Tomer Lange, Joseph (Seffi) Naor, and Gala Yadgar ({tlange, naor, gala}@cs.technion.ac.il), Technion—Israel Institute of Technology, Haifa, Israel

Parkview Hospital, Inc.
Full-Time User Experience Research Specialist II

Parkview Hospital, Inc. is seeking a full-time User Experience Research Specialist II in Fort Wayne, IN, to specialize in Human Computer Interaction (HCI) research including leading the design strategy of wireframes or interactive prototypes; collaborating with the Informatics research group to conduct research activities either in field or laboratory settings; and leading synthesis of qualitative and quantitative data to uncover user needs that result in actionable recommendations to drive user centered ideation and design goals. Send resume to Mark Webster, Talent Acquisition, at Mark.Webster@parkview.com.

[CONTINUED FROM P. 140] get two damaged packages you are done: container C has only good packages. But in the worst case, you will get one damaged package and one good one. Keep looking in the container, say B, having the good package. If you ever find a bad package, then, container C has only good packages. Otherwise, after examining the 501st package from B, you will know B has only good packages.

The trouble with the Question 2 strategy is approximately half the time, you will inspect at least 501 packages. It is very unlikely the Warm-Up strategy will require the inspection of even a fraction of that number.

Question 3: What is the likelihood the Warm-Up 1 strategy will require the inspection of more than 21 packages? An answer within a factor of two is fine.

Solution to Question 3: Approximately 1 in 1,000. In the first round, we get one bad and two good packages. We discard the container having the bad one and then have just containers B and C. Now, let's focus on the mixed container, say B. For B to survive round k, its package has to be good. The probability that this is so is ½ (slightly less because there have been only good packages prior to the k^{th} round). Thus, for B to survive until round 10 (corresponding to the first round where three packages are inspected and 9 rounds when two packages are inspected for a total of 21 packages), B must survive 10 tests with a ½ chance of getting caught, which is 1/1,024. So, the Warm-Up 1 strategy requires fewer inspections than the Question 2 strategy most of the time.

What is a strategy to get 1,000 good packages while minimizing the average case number of inspections?

Upstart 1: Suppose the mixed container has a fraction f of good packages and 1-f of bad and $f > 1$-f. For some value of f, could the Warm-Up 1 strategy require fewer inspections than the Question 2 strategy in the worst case? In the average case?

Upstart 2: What if all containers have different fractions of good packages. We know the fractions, but we do not know which container has which fraction. Also, there are at least 1,000 good packages among all containers. What is a strategy to get 1,000 good packages while minimizing the average case number of inspections?

Upstart 3: Same as Upstart 2 but we want 1,000 packages of which at most a fraction b could be bad.

Dennis Shasha (dennisshasha@yahoo.com) is a professor of computer science in the Computer Science Department of the Courant Institute at New York University, New York, NY, USA, as well as the chronicler of his good friend the omniheurist Dr. Ecco.

All are invited to submit their solutions to upstartpuzzles@cacm.acm.org; solutions to upstarts and discussion will be posted at http://cs.nyu.edu/cs/faculty/shasha/papers/cacmpuzzles.html

DOI:10.1145/3595917

Dennis Shasha

Upstart Puzzles
Sampling the Goods

Clever sampling for the worst/average case.

THE BASIC PRINCIPLE of sampling is to gain as much information as possible from each sample. If you know what you will find in a sample, then there is no point in sampling it. The more uncertainty, the more you learn—the basic principle of information theory.

Warm-up 1: Consider that you are faced with three containers, each containing 1,000 packages as in the figure. One container holds packages that are all damaged. One container holds packages that are all good. One container contains half good and half bad. If you want 1,000 good packages, describe an inspection strategy that guarantees you can identify 1,000 good packages and determine which container contains all good, which contains all damaged, and which is mixed. Note that inspection does not destroy a good package.

Solution to Warm-up 1: Call the containers A, B, and C. Take packages in turn from each container. If any package from a container is damaged, then ignore that container in the sequel. If you get to the point there is only one container left, that is the container you want and no more inspections are needed. This works because any container that has any damaged package is either mixed or fully damaged. The container without any damaged package must therefore be the good one. You will find the container all of whose packages are damaged in the first round robin for sure. If you never find the mixed one after choosing 500 from each of two containers, then one more examination will determine which container has all good packages and which container is mixed.

There are three cargo containers. Each container contains 1,000 packages. One contains packages that are all damaged. One contains packages that are all good. One contains 500 of each. You want 1,000 good packages. Inspecting a single package takes an hour, so you would like to inspect as few as possible. What strategy should you use and how long will you expect to inspect?

Warm-up 2: If you use the Warm-up 1 strategy, what is the smallest number of packages that need to be inspected for you to be sure to know which container contains only good packages?

Solution to Warm-up 2: Two packages is the minimum to inspect. If the packages from containers A and B are both bad, then surely C has only good packages.

Question 1: What is the maximum number of packages you might need to inspect using the Warm-up 1 strategy?

Inspection does not destroy a good package.

Solution to Question 1: The worst case for that strategy is the first round finds two good (undamaged) packages and one damaged one. Subsequent rounds on the remaining two containers keep finding good packages until the 500th package of the mixed container is selected. At that point, you have 1,000 good packages. So this requires 1,001 inspections.

Question 2: The Warm-Up 1 strategy requires many inspections in the worst case. Is there a strategy that guarantees you will not need more than 502 inspections to identify 1,000 good packages?

Solution to Question 2: Take one from each of two containers, say A and B. If you [CONTINUED ON P. 139]

IMAGE BY ANDRIJ BORYS ASSOCIATES. USING SHUTTERSTOCK

ACM BOOKS

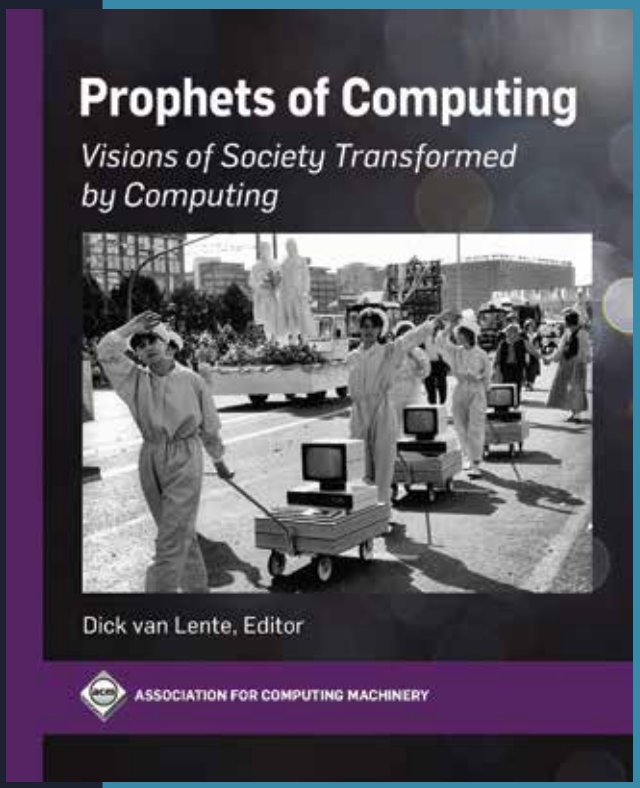
Collection II

When electronic digital computers first appeared after World War II, they appeared as a revolutionary force. Business management, the world of work, administrative life, the nation state, and soon enough everyday life were expected to change dramatically with these machines' use. Ever since, diverse prophecies of computing have continually emerged, through to the present day.

As computing spread beyond the US and UK, such prophecies emerged from strikingly different economic, political, and cultural conditions. This volume explores how these expectations differed, assesses unexpected commonalities, and suggests ways to understand the divergences and convergences.

This book examines thirteen countries, based on source material in ten different languages—the effort of an international team of scholars. In addition to analyses of debates, political changes, and popular speculations, we also show a wide range of pictorial representations of "the future with computers."

http://books.acm.org



Prophets of Computing
Visions of Society Transformed by Computing

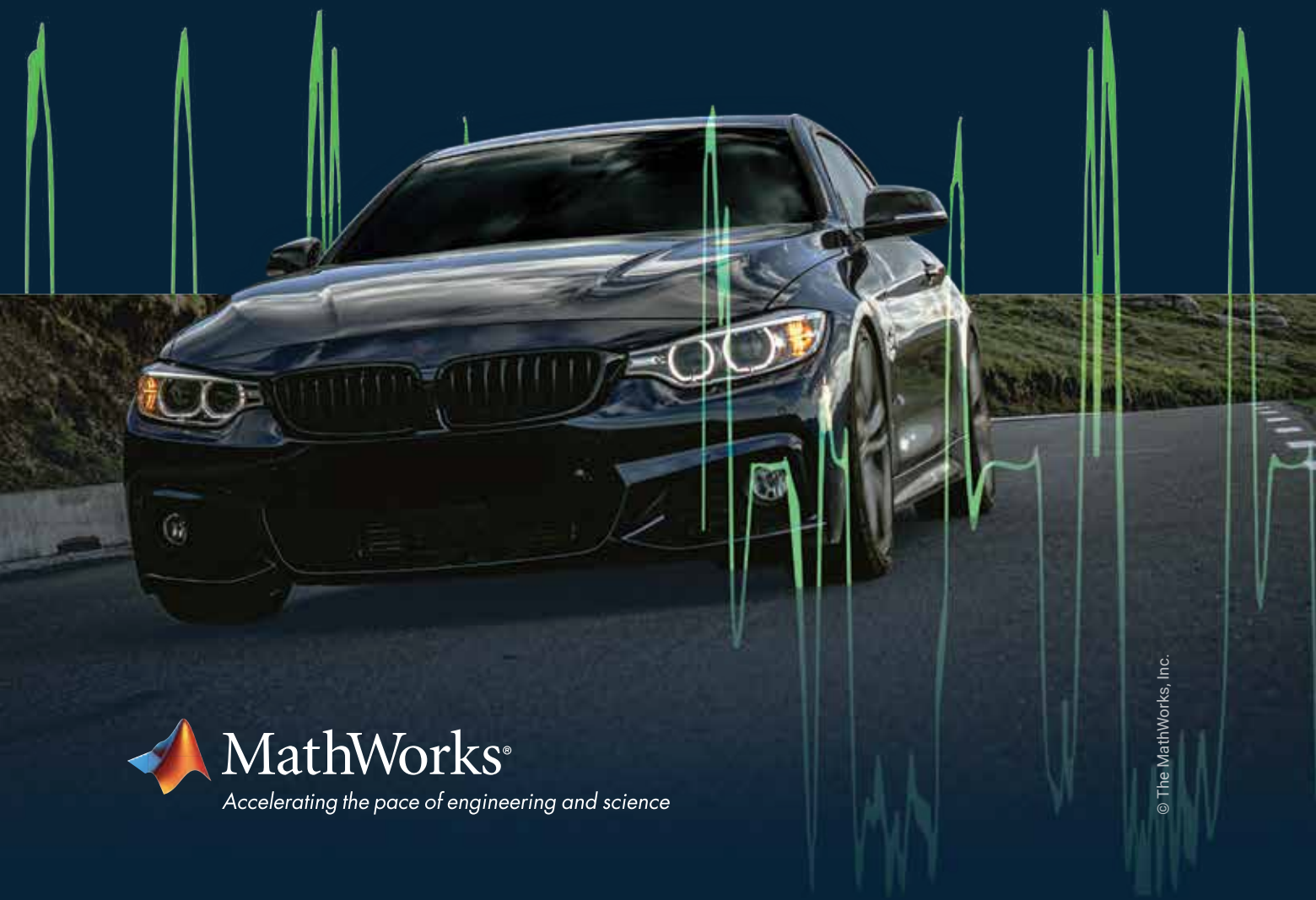
Dick van Lente, Editor
Erasmus University (retired)

ISBN: 9781450398152
DOI: 10.1145/3548585

MATLAB FOR AI

Boost system design and simulation with explainable and scalable AI. With MATLAB and Simulink, you can easily train and deploy AI models.

mathworks.com/ai



MathWorks®

Accelerating the pace of engineering and science