COMMUNICATIONS OF THE ACM

CACM.ACM.ORG

12/2020 VOL.63 NO.12

Silicon Politics

Measuring Internet Speed

XNOR-Networks

The Dark Triad and Insider Threats in Cyber Security

Association for
Computing Machinery

acm

ISBN: 978-1-970001-12-9 DOI: 10.1145/2792790
http://books.acm.org
http://www.morganclaypoolpublishers.com/vr

ECOOP / AARHUS 2021 / JULY 12-16

THE 35TH EDITION OF ECOOP

https://2021.ecoop.org

ECOOP is Europe's longest-standing annual Programming Languages conference, bringing together researchers, practitioners, and students to share their ideas and experiences in all topics related to programming languages, software development, object-oriented technologies, systems and applications. It welcomes high quality research papers relating to these fields in a broad sense; solicits both innovative and creative solutions to real problems as well as evaluations of existing solutions—evaluations that provide new insights; encourages the submission of reproduction studies

AREAS OF INTEREST

Design, implementation, optimization, analysis, and theory of programs, programming languages, and programming environments.

PUBLICATION

Affordable CC-BY open access in Dagstuhl LIPIcs
or as journal first in ACM TOPLAS or Elsevier SCP.

GENERAL CHAIR

Anders Møller, Aarhus University

PROGRAM CHAIR

Manu Sridharan, UC Riverside

PROGRAM COMMITTEE

Jonathan Aldrich
Walter Binder
Eric Bodden
Camil Demetrescu
Werner Dietl
Jens Dietrich
George Fourtounis
Colin Gordon
Michael Greenberg
David Grove
Murali Krishna Ramanathan
Burcu Kulahcioglu Ozkan
Viktor Kunčak
Julia Lawall
Mira Mezini
Todd Mytkowicz
Robert O'Callahan
Hakjoo Oh
Uday P. Khedker
Hila Peleg
Fernando M. Q. Pereira
Michael Pradel
Sukyoung Ryu
Alexandra Silva
Alexander J. Summers
Sam Tobin-Hochstadt
Omer Tripp
Eelco Visser
John Wickerson
Lingming Zhang
Lu Zhang
Elena Zucca

AARHUS UNIVERSITY

AITO

acm In-Cooperation

acm SIGPLAN

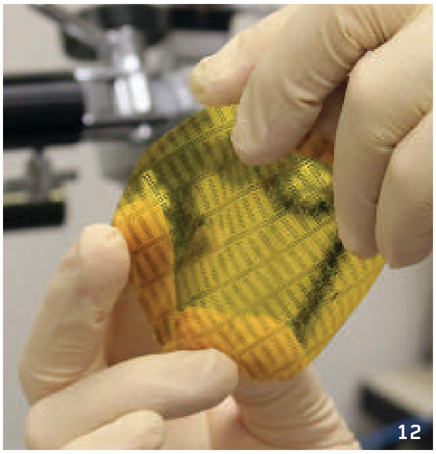
COMMUNICATIONS OF THE ACM

Departments

Last Byte

Association for Computing Machinery
Advancing Computing as a Science & Profession

News

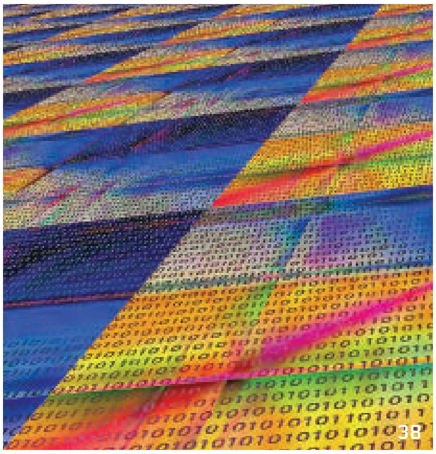

12

Viewpoints

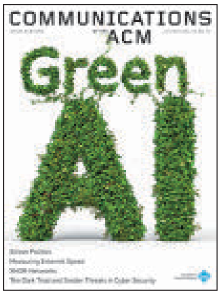
IMAGE COURTESY OF UNIVERSITY OF TEXAS AT ARLINGTON

12/2020
VOL. 63 NO. 12

Practice



Articles' development led by acmqueue
queue.acm.org



About the Cover:
Computational power for deep learning research takes deep pockets and leaves a big carbon footprint. This month's cover story (p. 54) advocates a practical solution for cutting both. Cover illustration by Lisa Sheehan.

LEFT IMAGE BY ALESANKO RODRIGUEZ; RIGHT IMAGE BY LISA SHEEHAN

Contributed Articles

Watch the authors discuss this work in the exclusive *Communications* video.
https://cacm.acm.org/videos/green-ai

Review Articles

Watch the authors discuss this work in the exclusive *Communications* video.
https://cacm.acm.org/videos/measuring-internet-speed

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.

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

WEBSITE
http://cacm.acm.org

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

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

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

Media Kit acmmediasales@acm.org

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

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

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

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

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

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

Printed in the USA.

Association for
Computing Machinery

PLEASE RECYCLE THIS MAGAZINE

DOI:10.1145/3429267

Vinton G. Cerf

Repairability Redux

I wrote about repairability in the February 2020 issue of *Communications* (p. 7) and here I am at year's end harping on the same topic. My excuse is COVID-19. I have been at home for the past six months while my normal schedule would have had me on the road three weeks out of four. Of course, like many of you, I have been all over the world—virtually—since mid-March. There are days when I can visit Australia and Austria and be home in time for dinner. So, what does that have to do with stuff that breaks? Mostly, I am actually here when it does. Under more normal conditions, my wife would have to call a repair person to fix or replace a broken item. Now that I am home, I am sometimes the one who discovers the problem, or I am told about it before a repair service gets the call. I am an engineer of sorts, so broken things attract my attention. Engineers love problems to solve. "Fix me! Fix me! You can do it!" Of course, if you are like me, you go to the hardware store three times: first to get the stuff you need, second to get the stuff you forgot, and third to get the stuff you need to fix what you broke. My basic rant is that manufactured goods today do not seem to take into account the possibility of repair.

Case in point: a broken soap dispenser. Expensive, shiny chrome built-in kitchen soap dispenser whose <insert adjectives> spring gave out. OK, no problem. Look up replacement parts on the Internet. Found it! Only $4.79 too. I ordered the part and it arrived in the mail a few days later. OK, just pull out the old one and slip in the new one … Hmm. The new siphon has a diameter of .29 inches. The hole is .25 inches. <Many bad words>. Why didn't they provide diameter specifications? Well, the original parts are 22 years old and I guess they changed sizes somewhere in between. Maybe I can sand down the .29-inch tube … Nope, can't do that without making the plastic tube too brittle. Time to call my "go-to engineer" friend who has a $30,000 Computer Numerical Control (CNC) machine that can pretty much mill anything.

This CNC thing has a bazillion interchangeable drills and other gadgets for making measurements accurate to a ten thousandth of an inch. The machine is programmable to carry out complex milling operations on a variety of materials including my chrome-plated brass soap dispenser fixture. Measuring carefully, we are going to drill out the .25-inch hole until it is .29 inches in diameter so the new siphon tube will fit snugly into the enlarged hole. Here's the right drill bit—#108—set for 6000 RPM. Holey Moley! The drill bit got yanked out of the chuck! My engineer friend says, "Brass grabs like that." OK, new plan. Let's get a much smaller drill bit and program the machine to ream out a larger hole by making a circular traverse multiple times, drilling deeper on each traverse. Sonfagun, that works! Habemus soap dispenser!

My basic rant is that manufactured goods today do not seem to take into account the possibility of repair.

In the last several weeks I have encountered several similar problems with "replacement parts" that don't quite fit the <ancient> piece of equipment I am trying to repair. I am beginning to appreciate Cuban ingenuity. Have you seen all those cars from the 1950s in Havana? I think manufacturers today should take lessons from the LEGO company. They have produced interchangeable LEGO parts since 1932. If you are going to make products that are intended to last for decades, you should maintain spare-part compatibility for the lifetime of the product. That's what standards are for. Maybe 3D printing is a partial solution for some products since printing the part on demand might be less expensive than maintaining inventory. I think my basic complaint is that if something is advertised as a *replacement part*, it should really be a replacement part that fits. Well, there is always eBay, I suppose. Thanks for listening. C

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

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

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

DOI:10.1145/3427786 http://cacm.acm.org/blogs/blog-cacm

Pitting Computers Against Each Other … in Chess

Guest blogger Monroe Newborn on the 50th anniversary of ACM computer chess tournaments.

Monroe Newborn
The Laughing is Over
September 18, 2020
https://bit.ly/35Nc1SS

For those of us involved in programming computers to play chess, it has been a great adventure. ACM annual tournaments began in 1970 (50 years ago!) and were hosted year after year for a quarter-century by the organization. They were terrific catalysts for progress in the field, and deserve major credit for the eventual 1997 defeat of then-World Champion Garry Kasparov.

I feel human intelligence has been vastly overrated. We humans haven't learned how not to fight wars over various explanations of how the universe or man came into being. The telescope that allows us to peer into the universe at 1,000,000,000,000,000,000,000 stars is only 400 years old. Women were burned alive as witches only 300 years ago. With the universe in existence for 14.7 billion years, the Earth in existence for 4.6 billion years, and humans in existence for several million years, there is an excellent chance at least one of those stars may now and/or in the past have supported some sort of intelligent life. I suspect there might be creatures far more advanced than Earthlings. One of the great disappointments of my life is that we have not made greater efforts to explore the universe, especially our Moon and Mars.

Historically, there have been arguments supporting the belief the animals of this planet do not possess the mental capabilities of humans. They were not thought to be able to use tools to solve problems, and they were not thought to be able to recognize themselves in a mirror. More and more, we find animals are more intelligent than they have been given credit for.

Researchers in the field decided in the 1940s that chess was the task over which human intelligence might be studied. The definition of intelligence, of course, is problematic. How do we say someone is more intelligent than someone else? Chess provides an environment to study this. Ratings are assigned to chess players based on their performance against opposition. If one person consistently defeats another person, the former will be higher-rated and can be considered more intelligent than the latter at playing chess. That, in part, was the motivation for using chess as a barometer of intelligence. There is also no luck involved, as with many other games.

This brings us back to the game of chess and, in particular, ACM's role in it. In 1970, ACM's Annual Conference took place at the New York Hilton. Ken King, head of Columbia University's computer center, served as co-chairman with me of the conference's Special Events program. About the same time, Tony Marsland, a Bell Telephone Laboratories researcher, approached me suggesting a demonstration of his chess-playing computer at the conference as a special event. We met, discussed this, and concluded we could put together something more exciting: the first chess tournament exclusively for computers. It was named the U.S. Computer Chess Championship. In addition, we included a computer music festival and a computer art festival in the program. These constituted our special events. I have forgotten how the music and art festivals turned out. I do recall Charles Dodge's music involved transforming the intensity of energy from the sun into musical notes, and the resulting music's randomness was surprisingly delightful.

Marsland and I rounded up six contestants for the chess tournament.

Hans Berliner, who had been the world correspondence chess champion, was a doctoral student at Carnegie Mellon University at that time; he entered his program J. Biit. Ken King arranged for it to run on Columbia University's powerful IBM 360/91 during the competition. Berliner's program entered the tournament as the favorite.

Northwestern University students David Slate, Larry Atkin, and Keith Gorlen entered their program CHESS 3.0. It ran on their university's CDC 6400, a powerful machine, but not in the same class as Columbia's IBM 360/91.

Marsland entered his program, The Marsland CP; it ran on a Burrough's B5500 located in Burrough's New York City sales office.

Running on IBM 360/65s were two other entries: COKO III, developed by Dennis Cooper and Ed Kozdrowicki at Bell Telephone Laboratories' Whippany facilities, and SCHACH, developed at Texas A&M by Franklin Ceruti and Rolf Smith, U.S. Air Force captains at the time.

Lastly, Chris Daly, working with Ken King (not Columbia's Ken King), brought their computer, an IDIOM system based on a Varian 620/i processor, to the site.

Three entrants used terminals connected to their remote computers, two others spoke by telephone to a human operator at their computers' sites, and one was at the site.

Jacques Dutka, a mathematician known for calculating the square root of 2 to a million decimal digits, served as tournament director.

Missing from the competition was Mac Hack, developed at MIT by Richard Greenblatt, established leading up to the tournament as the strongest chess-playing program. It had competed in a number of human tournaments and was rated around 1600, the level of a good high school player.

So began a very significant, long-lasting experiment. Could a computer be designed to exhibit the intelligence of a chess expert, master, grandmaster, or world champion? Could one be designed to match the chess intelligence of the top human mind? What would it take? How long would it take? In 1970 there were no cellphones, no email, no drones, no self-driving cars, no Siri. Yet, the computer revolution was heating up!

Grandmasters were generally in denial in 1970. Some contended good chess players used intuition when playing chess, and intuition could not be programmed. The programs were the laughingstocks of the top chess players. That was the situation when the first ACM U.S. Computer Chess Championship was held.

The format of the competition was a three-round Swiss-style tournament beginning Aug. 31, 1970, and ending Sept. 2, 1970. Entries had two hours to make their first 40 moves, then 30 minutes to make each successive 10 moves. Bugs cropped up as the competition went on, with the most dramatic seen in the early moves of the round 1 game between The Marsland CP (White) and J. Biit (Black). Marsland's program made the worst possible 8th and 9th moves, leading to a quick victory for J. Biit and laughter from the audience of computer and chess experts. Attendees Grandmaster Pal Benko and International Master Al Horowitz may have been among those laughing.

In the next round, J. Biit was defeated by the Slate/Atkin/Gorlen program, with the audience cheering CHESS 3.0's 47th move, a short-term sacrifice leading to an easier, shorter path to victory. Quite unlike a human tournament, the audience was very vocal as the games progressed, cheering and laughing. Also unlike human tournaments, programmers would get together for coffee after the games ended, to discuss the day. A close community of programmers developed over the years.

CHESS 3.0 went on to win the tournament, winning all three games. It dominated the field for a decade, until Ken Thompson's BELLE arrived in the late 1970s with special chess hardware. BELLE stayed on top until Bob Hyatt's CRAY BLITZ and Hans Berliner's HITECH caught up in the middle 1980s. In the late 1980s and early 1990s, IBM's Deep Blue, developed by Feng-Hsiung Hsu with major help from Murray Campbell, Joe Hoane, and Jerry Brody, with Chung-Jen Tan serving as boss, rose to the top of the pack. Deep Blue, playing grandmaster-level chess, went on to defeat Garry Kasparov in their classic 1997 match, a landmark in the world of artificial intelligence.

(Tan and I go back to the late 1960s, when he was a doctoral student at Columbia University and I was a young professor there. We lived in the same apartment building. We published several papers together in the field of automata theory; most prominently, one in the January 1970 issue of *IEEE Transactions on Computers* entitled Iteratively Realized Sequential Circuits.[1])

A half-century later, chess programs are so much better than top humans that there is no contest. The top players use computers to help them learn to play better! The laughing has ended.

Computers are, in round numbers, 1,000 times faster than in 1970, and their memories are clearly more than 1,000 times larger. Imagine driving a car that goes 1,000 times faster than the one you currently drive; a 10-mile drive to work at 60 mph would take 10 minutes, while at 60,000 mph it would take less than a second. Chess programs require, for each additional level of search, approximately four times the amount of time; a speedup of 1,000 allows computers to search about five levels deeper (4x4x4x4x4 = 1024). On top of faster computers and much larger memories, there have been many software improvements and even different approaches all together, such as the use of Monte Carlo search.

Over the years, ACM headquarters supported the yearly tournaments. I would like to single out two individuals in particular: Jim Adams and Joe DiBlasi. In addition, Drexel University professor Frank Friedman provided support, as did Ben Mittman, head of Northwestern University's Vogelback Computer Center. Lastly, British International Chess Master David Levy helped, especially in setting up the 1996 Kasparov versus DEEP BLUE match.

Reference

1. Arnold, T.F., Tan, C.J., and Newborn, M. Iteratively realized sequential circuits. *IEEE Trans on Computers* (Jan. 1970), 54–66.

My books, one with David Levy, chronicle the great progress computers have made at chess and ACM's involvement in it:

Computer Chess, Academic Press, New York, 1975.

How Computers Play Chess, with David Levy, W. H. Freeman, 1991.

Kasparov versus DEEP BLUE: Computer Chess Comes of Age, Springer-Verlag, 1997.

Deep Blue: An Artificial Intelligence Milestone, Springer, 2002.

Beyond Deep Blue: Chess in the Stratosphere, Springer, 2011.

Monroe "Monty" Newborn, formerly chairman of ACM's computer chess committee, was a professor of electrical engineering in Columbia University, and later a professor of computer science in McGill University, where he is currently a Professor Emeritus.

ACM Transactions on
Computing for Healthcare
2020
health.acm.org

news

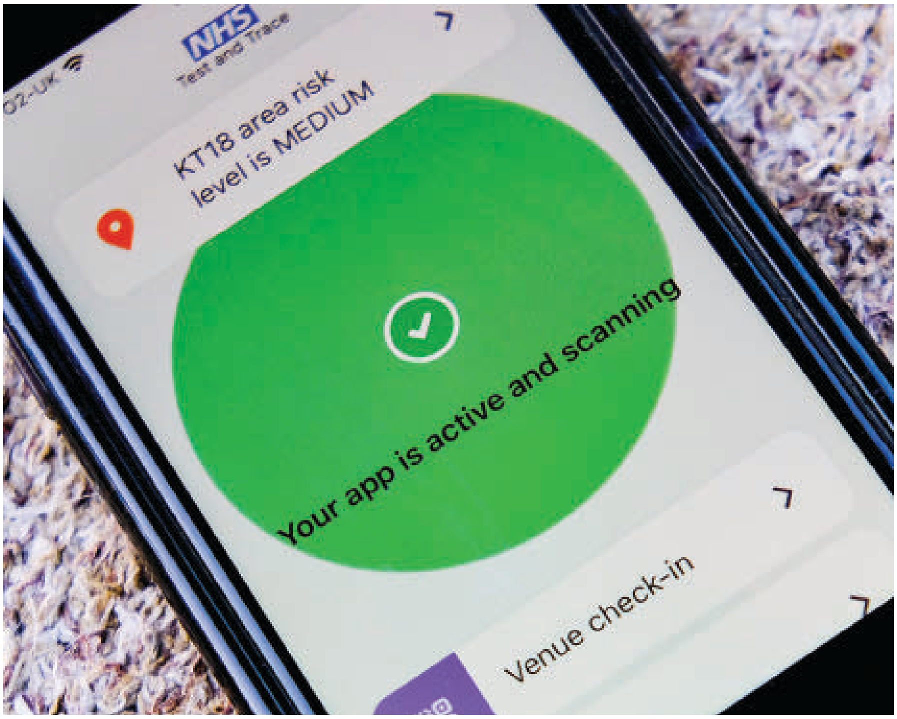
Science | DOI:10.1145/3427933 Neil Savage

Tracking COVID, Discreetly

Tracing the contacts of those who come into contact with the coronavirus is not that simple.

As the world continues to grapple with the coronavirus pandemic, health officials are relying on a tried-and-true method of limiting the spread of the potentially deadly disease: contact tracing. Figuring out who has been close enough to an infected person long enough to catch the disease, then taking steps to prevent those people from passing it to others, is a method that dates back to the 1920s, when health authorities used it to rein in the spread of syphilis. In the era of smartphones, it seems only natural to add a technological dimension to contact tracing.

Using smartphone apps for contact tracing raises questions, though. For one thing, it is not entirely clear how effective that is; the answer depends on both how well a smartphone can measure contacts and on how many people actually decide to use the apps. Perhaps the chief concern, though, is privacy. How do you design a system that identifies who has been in contact with whom without giving all sorts of personal information to governments or data thieves that might abuse it?



O2-UK
NHS
Test and Trace
KT18 area risk level is MEDIUM
Your app is active and scanning
Venue check-in

IMAGE BY RICHARD M LEE/SHUTTERSTOCK.COM

Civil libertarians have raised alarms about potential invasions of privacy. A report in June from Amnesty International warned governments' collection and storage of too much information about individuals posed a significant threat, especially in some countries. "Bahrain, Kuwait, and Norway have run roughshod over people's privacy, with highly invasive surveillance tools which go far beyond what is justified in efforts to tackle COVID-19," said Claudio Guarnieri, head of Amnesty International's Security Lab, in a statement that accompanied those findings. The Norwegian government, also facing pressure from the European Data Protection Authority, suspended use of its app, called Smittestopp.

Problems with these apps included identifiers that could be tied to people. Smittestopp made users register with their telephone number, for instance, while Bahrain, Kuwait, and Qatar required use of a national identification number. Early apps often had location-based tracking, which can make it possible to re-identify anonymized users and keep tabs on potentially sensitive information such as who someone was visiting or whether they were taking part in protests. Several countries, including France, Iceland, and Singapore, had centralized storage of the data, where access to it would be out of an individual's control.

The Tech Giants

Over the summer, however, governments adopting apps moved to a different model, and many now rely on the Exposure Notification System specification jointly developed by Apple and Google. In that system, individual phones generate random numbers, which change every few minutes, and share them with nearby phones over Bluetooth. The Bluetooth signal can be used to estimate the distance between phones and the length of the contact—being within six feet of someone for 15 minutes is generally defined as a contact. The data is stored on the user's phone and automatically deleted after a specified period of time.

An app avoids the risk of someone, whether the government or a hacker, getting their hands on personal data by not collecting such data in the first place.

If someone using the app tests positive for COVID-19, they enter that into the app, which then uploads all stored contacts to a master list on cloud storage platforms. Another user's app periodically checks that master list, and if it finds its own key, it notifies the user that she should get tested or self-quarantine. Once the incubation period has passed, keys are deleted from the master list. The system specifically bars apps from collecting GPS data.

An app avoids the risk of someone, whether the government or a hacker, getting their hands on personal data by not collecting such data in the first place. "A solution should not be about confidentiality or anonymity. It should be about privacy, in which even the company doesn't know who you are," says Ramesh Raskar, professor in the Massachusetts Institute of Technology's Media Lab who launched the PathCheck Foundation. The non-profit has developed an app based on the Apple-Google framework to track COVID-19 cases using "privacy by computation." Data can only be seen by someone with physical access to the phone handset, and even then the amount of useful information that could be obtained would be limited, Raskar says.

Location, Location, Location

Last March, Robert Kleinman, a psychiatrist then at Stanford University and now at Massachusetts General Hospital, decided to combine his interests in geospatial data and access to healthcare. He and software engineer Colin Merkel designed a prototype tracking app that used GPS location data to identify where exposures took place. They eventually scrapped it, as did most other developers who focused on GPS early on.

The problem, aside from possible public discomfort with having one's location tracked, is that GPS location is not precise enough for contact tracing.

Milestones

ACM Releases Study of U.S. Bachelor's Programs in Computing

ACM recently released its eighth annual Study of Non-Doctoral Granting Departments in Computing (NDC), with the aim of providing a comprehensive look at computing education.

This year's ACM NDC study includes enrollment and degree completion data from the National Student Clearinghouse Research Center (NSC). In previous years, ACM directly surveyed computer science departments, and would work with a sample of approximately 18,000 students. By accessing the NSC's data, the ACM NDC study now includes information on approximately 300,000 students across the U.S., allowing for a more reliable understanding of the state of enrollment and graduation in bachelor's programs.

The study also includes data from private, for-profit institutions.

Important findings of the study include:

- **Between the 2017/2018 and the 2018/2019 academic years, there was a 4.7% increase in degree production across all computing disciplines, with the greatest increases in software engineering (9%) and computer science (7.5%).**
- **The representation of women in information systems (24.5% of degree earners in the 2018/2019 academic year) and information technology (21.5% of degree earners in the 2018/2019 academic year) is much higher than in areas such as computer engineering (12.2% of degree earners in the 2018/2019 academic year).**
- **Bachelor's programs had a stronger representation of African American and Hispanic students than Ph.D. programs, as recorded by the Computer Research Association's (CRA) Taulbee Survey.**
- **In some disciplines of computing, African Americans and Hispanics are actually overrepresented, based on their percentage of the U.S. population.**
- **Based on aggregate salary data from 89 non-doctoral-granting computer science departments (including public and private institutions), the average median salary for a full professor was $109,424.**
- **Of 40 non-doctoral-granting departments reporting over 56 faculty departures, only 10.7% of faculty departed for non-academic positions. Most departed due to retirement (46.4%) or other academic positions (26.9%).**

The study was authored by Stuart Zweben, professor emeritus, Ohio State University; Jodi Tims, a professor at Northeastern University; and ACM education and professional development manager Yan Timanovsky. By employing NSC data in future studies, the co-authors said they were confident an even fuller picture will emerge regarding student retention with respect to computing disciplines, gender, and ethnicity.

"It's quite remarkable for a consumer technology to get within three meters or five meters, but there is a lot of variability in that, and it's different in indoor settings and outdoor locations," Kleinman says. "It would just have a lot of limitations for identifying contacts in a reliable way, and you would end up getting a lot of false positive and false negative identifications."

Either is a problem. False negatives could mean missing actual cases of disease transmission, but false positives would result in people being told to isolate themselves unnecessarily, which can be disruptive to work, personal life, and mental health.

There are also problems with Bluetooth, which is not designed to measure distance. The space between phones can be inferred by the strength of the signal, but the orientation of the phone, a wall, or even the user's own body can alter apparent signal strength.

Keeping Score

One way to deal with that weakness in Bluetooth is to use scoring algorithms to help decide whether a phone contact is enough of an in-person contact to trigger an alert, says Stefano Tessaro, a cryptographer and computer security expert at the University of Washington (UW). Tessaro and a loose coalition of researchers from UW, Microsoft Research, the University of Pennsylvania, and the Boston Public Health Commission developed what they dubbed PACT, privacy-sensitive protocols and mechanisms for mobile contact tracing.

It would be useful to come up with formulas that use factors such as signal strength and length of contact to score whether something counts as actual exposure, rather than triggering an alert for, say, every student who walks by a professor's window and later tests positive. The difficulty, Tessaro says, is that despite a large number of cases, the disease is still rare enough that real-world data is lacking. "There's not enough positive cases, fortunately, to be in a situation where you can really see a lot of such false positives," Tessaro says. Additionally, the same restrictions that protect users' privacy also make it more difficult for researchers to collect data that can tell them how good a job an app is doing at correctly identifying contacts.

One outstanding question is how many people must adopt contract tracing apps for them to be effective in slowing the pandemic.

While Tessaro understands why people might be uncomfortable having their location tracked, he also recognizes public health experts would love to have GPS data help them trace the spread of the disease and identify hotspots. He and his colleagues have proposed what they call narrowcasting, in which a user's phone collects its own location data but does not send it to anyone. Then, if a health department finds an infected person was in a particular park or grocery store at a given time, it could broadcast that information through an app, and if the health department information matches data stored on the phone, the user gets an alert.

Of course, old-fashioned manual contact tracing does not strictly protect people's privacy, either. Health workers talk with infected people and ask where they have been and who has been near them. "It's considered a fairly essential public health approach to addressing infectious disease. The question of where the appropriate line is, that's really a question for society," Kleinman says. One major difference is that traditional contact tracing starts with a known infected person and builds outward, ignoring those who have not been in contact with a patient, whereas apps collect some amount of information from everyone who uses them, he says.

One outstanding question is how many people must adopt the contact tracing apps for them to be effective in slowing the pandemic, and uptake depends in part on how comfortable people are that the apps are safe to use. Models made in April by the Big Data Group at Oxford University in the U.K. suggested if 60% of the population would use them, it could stop the disease in its tracks. More recent pilot studies Oxford ran on England's Isle of Wight indicated the spread could be slowed if just 15% to 20% of people used the app as recommended. That includes scanning QR codes to check into stores and restaurants, so health authorities can keep tabs on those businesses.

Tessaro says a lot is still unknown about what impact contact tracing apps could have, but they are unlikely to provide a quick fix. He thinks they may work best as a supplement to human-run contact tracing, with the apps filling in information that would be difficult for people to find, such as who shared a subway car with an infected person. "The metric that people apply to these tools is that of a silver bullet, that there's one thing that is going to fix everything," he says. "But it's really not true."

Further Reading

Kleinman, R. and Merkel, C.
Digital Contact Tracing for COVID-19, *CMAJ*, 192 (24) 2020. www.cmaj.ca/content/192/24/E653

Chan, J., Foster, D., Gollakota, S., Horvitz, E., Jaeger, J. Kakade, S., Kohno, T., Langford, J., Larson, J., Sharma, P., Singanamalla, S., Sunshine, J., and Tessaro, S.
PACT: Privacy-Sensitive Protocols and Mechanisms for Mobile Contact Tracing, ArXiv, 2020. arxiv.org/abs/2004.03544

Singh, P., Singh, A., Cojocaru, G., Vepakomma, P., and Raskar, R.
PPContactTracing: A Privacy-Preserving Contact Tracing Protocol for COVID-19 Pandemic, ArXiv, 2020. arxiv.org/abs/2008.06648

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., and Fraser, C.
Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing, *Science*, Vol. 368, Issue 6491, 2020. science.sciencemag.org/content/368/6491/eabb6936

Exposure Notifications System: Helping Health Authorities fight COVID-19 https://www.youtube.com/watch?v=1Cz2Xzm6knM&feature=emb_logo

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

Technology | DOI:10.1145/3427942

Esther Shein

Softening Up Robots

Giving robots soft, artificial skin would enable them to work more closely with people.

When you picture a robot, you likely envision one large and rigid, with limited movement and an outer shell that is hard to the touch. Several projects currently underway seek to change that, with the use of soft, more human-like artificial skin.

Artificial skins include any surface-based device or distributed network of sensors that enable an agent to perceive mechanical deformations, touch, temperature, vibration, and/or pain, according to Ryan Truby, a post-doctoral fellow in the Massachusetts Institute of Technology (MIT) Computer Science & Artificial Intelligence Lab (CSAIL). Engineers are working to create skins that include as many of these sensations as possible, while also possessing high sensitivity and spatial resolution in sensing, he adds.

"Artificial skin will liberate today's robots and bring them to a higher level of consciousness," notes Yiannis Aloimonos, a professor at the University of Maryland who leads the school's Computer Vision Laboratory (CVL). "They will become much safer, especially when operating near people."

Robotic skin cells, in one sense, can do more than animal skin cells because they can sense proximity to a surface for anticipating and avoiding accidents, Aloimonos says. "Having this additional sense, the robots will be able to perceive their environment in greater detail."

The fusion of vision and tactility will make possible many new tasks that were not possible before, Aloimonos says. One problem being studied in his lab is how to teach robots to use tools. When humans use tools, we remember the feeling of a tool in our hands, and use this information to learn how to control its movement. "Recognition from vision and touch also becomes a new powerful technique,'' he says. "In general, any tasks where the robot comes into contact with something can be done better."

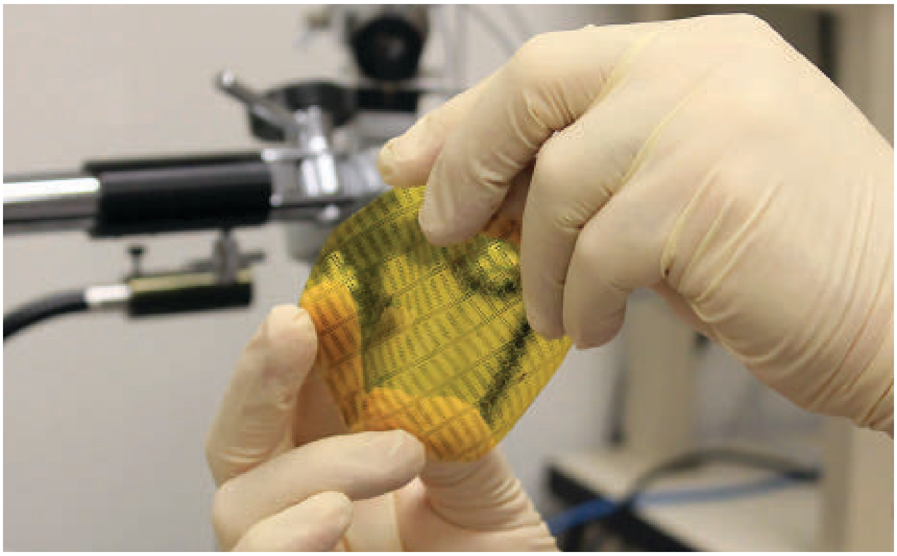

University of Texas at Arlington's patented smart skin technology, which the university says will give robots more sensitive tactile feeling than humans.

IMAGE COURTESY OF UNIVERSITY OF TEXAS AT ARLINGTON

Robots today are not able to help in dire situations like natural disasters. "There is a conspicuous absence of robots whenever we are in trouble,'' including in the midst of the coronavirus pandemic, says Aloimonos.

That is not to say the work robots do has not been important; it is just limited. Up until now, industrial and commercial applications of robotics have been large in scale and focused on factories, "and they are incredibly well-engineered for what they need to do, but they are not designed for use with humans," says Carmel Majidi, director of the Soft Machines Lab and an associate professor of mechanical engineering at Carnegie Mellon University (CMU).

One of the main ideas for developing soft robotics is to create machines that are safe for physical interaction with humans and won't injure us in a collision, Majidi says.

Often, industrial robots must be caged behind metal fences to prevent them from harming humans who unwittingly enter their areas of operation, says Majidi, who is spearheading an effort to develop new classes of soft materials in the fields of soft matter engineering and soft robotics.

A significant step in making machines safe to engage in physical contact with humans is to build them from materials that share lot of the same properties as us: water, fluids, soft materials, and nervous tissues that are soft and stretchy, Majidi says.

"Tactile sensing can be thought of as an 'artificial skin' technology that enables robots to have a sense of touch,'' he says. "They are used in a wide range of robotics applications, from humanoid robots to wearable technologies. They can also be used in soft robotics, but this is not their primary goal."

Autonomy in robotics can be advanced with improved perception, Truby says. "The abilities to sense touch, stretch, temperature, vibration, and many other forms of tactile and proprioceptive feedback are critical for the autonomy of any robot tasked with operating in the real world."

Artificial robot skins are a straightforward way to begin introducing these perceptual capabilities in robots, says Truby. "Innovations in robotic skins will likely be key to advancing the autonomy of robots designed for real-world deployment."

Additionally, he says, softer skins will improve the safety of human-robot interactions and also open up new ways for us to interact with machines.

These advances would take humans out of harm's way, he says. The field of soft robotics has generally shown that soft materials like the rubbers and hydrogels soft robots are being constructed from "do not just make robots safer for human-robot interactions; they also offer important advantages over materials like metals and thermoplastics that are used in 'rigid' robots today," Truby says.

These advantages can help soft robots become more resilient in harsh environments such as in a forest fire, for example, compared to their rigid counterparts, he says. "Thermoplastics would melt or burn from the heat of fires. "Excellent thermal conductors like metals can reach high temperatures in a fire, potentially damaging other onboard components due to heating. In contrast, soft robotic materials avoid these issues because they are often good thermal insulators and can demonstrate higher flame resistance."

Truby and his colleagues at MIT developed a system of soft sensors that cover a robot's body to enable its awareness of motion and position. "If we want to have soft-skin robots in the real world, we need to bring skins and sensing capabilities to these systems that can provide the feedback needed to autonomously control the robot's motion,'' Truby says. "We are going to have to solve the material engineering challenge of bringing those capabilities into soft robot bodies."

The research team tested its system on a soft robotic arm equipped with very simple sensors to predict its own position as it autonomously swings. The test proved eye-opening: "The very simple, inexpensive sensors we gave this soft robot arm enabled a deep neural network to provide a satisfactory estimate of the arm's 3D configuration as it moved about,'' Truby says. "To our knowledge, this is a first for the field of soft robotics."

He says the team is always working to improve its soft robots' capabilities and is now focused on improving the accuracy of the configuration estimations through new sensor skin designs and deep learning methods. "We are also working to use this feedback to dynamically control the arm for a variety of applications."

"We are working to improve the materials that go into these skins to enable tactile sensing, as we would expect from a sophisticated soft robot skin."

There are many types of artificial skins, but many are expensive to make, not easy to interface with soft robots, and are not scalable, Truby says. "We're fighting against cost and scalability,'' and accessibility. "The advantage of soft robots is they are cheap."

The MIT researchers worked with "commodity conductive silicone rubbers available right off the shelf," he says. Comparable skins comprised of other types of conductive rubbers, or materials like liquid metals or ionically conductive gels, have been made through relatively simple and inexpensive techniques such as 3D printing and molding, he says.

Ordinary rubbers is a good insulator but not a good conductor, Truby says. The materials used by the researchers "have conductive particles in them, so we can conduct electricity and can create sensory skins from them."

More complicated skins that can perceive multiple sensations and/or possess a higher density of "receptors" for these stimuli are often made with slower, more expensive microfabrication techniques, Truby says. "While microfabrication methods open up opportunities for more readily building with materials like thin metal films and conductive polymers, it's often very expensive to scale these techniques to produce artificial skins of large area,'' he says. "Thus, when designing artificial skins, there is a clear trade-off between performance, size or area, and cost."

Yet more expensive artificial skins can provide "a very high resolution of tactile and touch feedback,'' he notes. They provide a lot more functionality and are commonly used in prosthetics, Truby says. "But these available methods are expensive to implement, do not scale well for large soft robots like ours, and can be difficult to interface with soft robots,'' he says. "Our approach represents a new strategy for rapidly, inexpensively, and easily giving soft robots important perception capabilities, which remains a challenge in soft robotics."

Right now, MIT's soft robotic skins are not appropriate for tactile or touch sensing, Truby says. "We are working to improve the materials that go into these skins to enable tactile sensing, as we would expect from a sophisticated soft robot skin. We're also working on algorithms that complement these simple soft sensor skins and enable new contact sensing motifs from external cameras and actuation inputs."

Majidi's group at CMU is researching materials and systems that combine mechanical, electrical, and thermal properties and can function as artificial skin, nervous tissue, and muscle for soft robotics and wearables.

His lab created a self-healing material that repairs itself under extreme mechanical damage, according to CMU. The soft-matter composite material is composed of liquid metal droplets suspended in a soft elastomer; when damaged, the droplets rupture to form new connections with neighboring droplets and reroute electrical signals without interruption. Circuits produced with conductive traces of this material remain fully and continuously operational even when severed, punctured, or with material removed. The goal is for electrical connections to spontaneously restore themselves when damaged, Majidi says.

Applications for the soft material include bio-inspired robotics, human-machine interaction, and wearable computing. Because the material also exhibits high electrical conductivity that does not change when stretched, it could be used in power and data transmission, according to CMU.

The greatest challenge is delivering adequate power to electronics within these skins, Majidi says. Skin that is soft and stretchable cannot be charged with a “big hulking battery … so there’s the challenge of creating batteries that are soft and flexible,’’ which is one project his lab is working on.

The amount of energy contained in a small lithium battery is limited, he says, and if you want to incorporate sensors and electronics in artificial skin, today they are not capable of detecting their environment or physiology. Current lithium battery cells have enough power to operate the sensors that detect the environment or physiology, he says. “However, they don’t have enough power to also operate the microprocessor used to process the signals and the radio transceiver used to transmit the signals wirelessly. For there to be enough power for long-term sensing, signal processing, and wireless data transmission, the current miniature battery would have to be replaced with a much larger battery.”

A larger battery would add to the rigidity of the soft artificial skin technology, Majidi says. This introduces a challenge with supplying power to an artificial skin.

There are two potential solutions. The first is to use a battery that is soft, flexible, and stretchable, he says. This type of battery could be larger in size than the existing miniature lithium batteries and therefore supply more power, Majidi says. Moreover, because it is soft and compliant, it will not add to the rigidity of the soft artificial skin.

Another solution is to use energy-harvesting to convert energy sources in the environment into electricity. “My lab is currently working on harvesting energy from body heat using soft and stretchable thermoelectric generators, as well as from knee and elbow motions using a soft conductive material for triboelectric energy harvesting.”

Eventually, if enough energy can be harvested, “It could power motors to get robots to move and actuate,’’ he says. “That’s the really critical step.”

The inability of artificial skin to sweat is another challenge. Soft robots produce heat, and it is important to have materials that can manage that heat, Majidi says. “So we’ve been creating soft thermally conductive rubber that can help manage or dissipate the heat that’s generated by these machines and electronics.”

“We’ve been creating soft thermally conductive rubber that can help manage or dissipate the heat that’s generated by these machines and electronics.”

Sweating and synthetic skin are also being studied at Cornell University, where scientists have created a soft robotic muscle that stays cool and mimics how a hand moves.

The ability for a muscle to stretch is helpful in adapting to unpredictable circumstances, like walking through a disaster zone with rubble everywhere, says Cornell associate professor Rob Shepherd. “It’s hard to predict what the next meter of terrain will look like.”

Shepherd and other Cornell engineering professors use polymers—in this case, rubber—for the skin, since it stretches and is similar to human skin, he says. Like Truby, he adds, “It is also pretty cheap, and you can make them self-heal.’’

Such benefits come at many costs, Shepherd notes, one being that polymers and rubber don’t conduct heat very well, unlike metal. “Cooling robots hasn’t been as much of a problem as it will be in future for these polymer ones,’’ he says. “If we expect them to do a lot of work quickly, they’ll generate heat and that will damage the robot, so we have to dissipate the heat somehow.”

The Cornell researchers developed a system that automatically dilates pores in an actuator when the temperature reaches 86 degrees Fahrenheit. The idea is when the pores dilate, it allows for an increase in the rate of evaporation and cools the muscle’s body, similar to perspiration in hot-blooded animals, according to Anand Kumar Mishra, a post-doctoral student at Cornell working on the project. The job of the fluids is to actuate the muscle, says Mishra, “But when the robotic body temperature elevates, then the actuating fluid can also be used for perspiration to regulate the body temperature.”

This will enable a robot to operate for long periods without overheating, Shepherd said.

The idea of an artificial skin for robots has been researched for decades, and there are numerous applications being developed with human-robot collaboration (HRC), says Werner Kraus, head of the Robot and Assistive Systems department at the Fraunhofer Institute for Manufacturing Engineering and Automation IPA, in Stuttgart, Germany. “Artificial skin could enable this type of [HRC] interaction, since its sensors can detect a contact between human and robot and stop the robot’s movement according to safety requirements.”

Artificial skin offers more sensor data to perceive the environment, Kraus says. “In times of powerful machine learning algorithms, this data can be the enabler for implementing software on robots that makes them more autonomous and more efficient.”

Further Reading

Liu, D., Su, L., Liao, J., Reeja-Jayan, B., and Majidi, C.
“Rechargeable Soft-Matter EGaIn-Mn O2 Battery for Stretchable Electronics,” *Advanced Energy Materials*, Wiley Online Library, https://onlinelibrary.wiley.com/doi/abs/10.1002/aenm.201902798

Carmel Majidi: Self-Healing Electrical Material, https://www.youtube.com/watch?v=N_ijvkl51LM

inSPIREd – Dr. Ryan Truby - Can soft materials revolutionize the meaning of robotics?, https://www.youtube.com/watch?v=eRTyWgTHQZI

Chortos, A., Liu, J., and Bao, Z.
“Pursuing prosthetic electronic skin,” *Nature Materials*, https://www.nature.com/articles/nmat4671

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

Society | DOI:10.1145/3427936

Logan Kugler

Technologies for the Visually Impaired

The last decade has seen major advancements in technology for the blind and visually impaired, but problems remain.

Thanks to recent advances in technology, the blind and visually impaired are now able to lead more independent lives than ever.

The WeWALK Smart Cane is a great example of what is now possible. The WeWALK looks similar to the cane that some blind and visually impaired people have used for decades to avoid obstacles while walking, but it incorporates a few modern twists.

With a standard cane, you can still run into obstacles that are not immediately underfoot, like poles, tree branches, and barriers. The WeWALK, however, detects objects above chest level and audibly alerts you if you're getting too close, which can save you from a painful fall.

Also, when using a standard cane, you have to hold a smartphone in one hand to listen to directions, making it even more difficult and dangerous to navigate your environment. The WeWALK integrates with a smartphone's map app to read directions out loud, allowing you to keep one hand free.

Finally, the WeWALK costs less than $500, making it affordable for many of the estimated 10 million visually impaired people in the U.S., and 250 million worldwide.

The WeWALK is just one example of a larger trend. Technology to help the blind and visually impaired has become dramatically more powerful and significantly cheaper in the last 10 years. In the process, it has revolutionized how blind and visually impaired individuals navigate the world.

However, despite the progress, there are still major unresolved problems.

Technologies like sophisticated smartphone apps and artificial intelligence have empowered millions of blind and visually impaired individuals, but many websites are still inaccessible to the visually impaired, making it difficult or impossible to use key online services, tools, and experiences that the sighted may take for granted.

The WeWALK Smart Cane detects objects in the vicinity and provides audible alerts to prevent collisions.

IMAGE COURTESY OF WEWALK/WEWALK.IO

The result is an imbalance in assistive technology progress. On the one hand, it's never been easier or more affordable to acquire new, powerful accessibility technology; on the other, there are still serious accessibility issues with the technology that powers much of modern life.

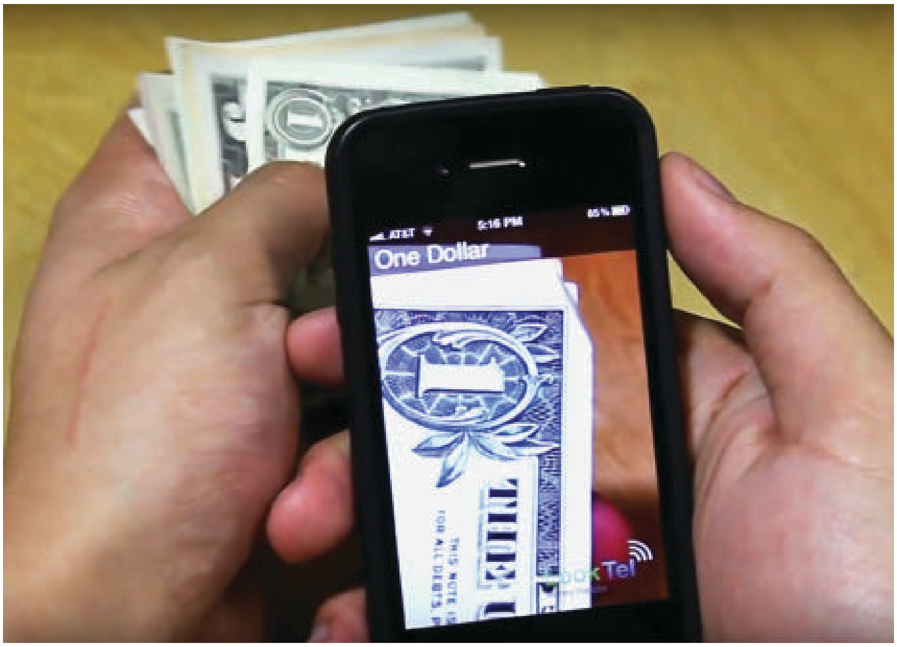
A Decade of Progress

Technological progress during the last decade has created major benefits for society as a whole, as well as for the blind and visually impaired in particular.

"Computer vision, fast mobile processors, high-speed wireless Internet, and cloud computing have enabled some truly fascinating innovations for resolving barriers in everyday life," says Aaron Steinfeld, associate research professor in The Robotics Institute at Carnegie Mellon University.

Most of today's innovations in technology for the blind and visually impaired are delivered via smartphone. These include sophisticated applications that empower the visually impaired to navigate, count money, and obtain assistance with basic tasks. They also include tools such as artificial intelligence and machine learning to recognize facial expressions and describe the external world in real time.

The sheer ubiquity of smartphones has made them the perfect enabling device for the blind and visually impaired. The percentage of U.S. adults who own smartphones has risen from 35% in 2011 to 81% in 2019, according to Pew Research, and Cisco estimates more than 70% of the global population will have mobile connectivity by 2023.



The LookTel Money Reader app captures an image of paper money on a smartphone, and speaks aloud the denomination of the bill.

IMAGE COURTESY OF LOOKTEL/YOUTUBE.COM

Navigation is a huge part of the value smartphones provide for the blind and visually impaired.

GPS-enabled devices mean you now have access to accurate, audible real-time navigation pretty much wherever you go. At least one popular navigation app—Google Maps—offers detailed voice guidance for the blind and visually impaired.

Smartphone navigation integrations also matter. The WeWALK Smart Cane's app integrates with the actual cane hardware, so you receive turn-by-turn instructions and clockwise navigation directions right from the cane itself, without having to remove your smartphone from your pocket.

Navigation isn't limited to walking down the street, either; smartphone applications now make it easier for the blind and visually impaired to navigate other difficult situations. For example,

- LookTel Money Reader helps the visually impaired count money. Usesr capture an image a picture of a bill on their smartphones, and the app speaks the denomination of the bill.
- TalkingTag LV is an iPhone app that allows users to label their surroundings with special coded stickers. When scanned by the app, the stickers play an audio message recorded by the user, enabling them to custom-label their environment. (The stickers only work when using the same phone, however; there is no database of locations labeled by others for users to access.)
- The Aira app connects the blind and visually impaired to remote agents who assist with navigation, reading printed media, and other everyday tasks. The app is free to download and use.

In all cases, the value of these apps is that they're affordable, a big change from a decade ago, says Jeffrey Bigham, an associate professor in the Human-Computer Interaction and Language Technologies Institutes at Carnegie Mellon University, who works on accessibility issues. "When I did my Ph.D. in the 00s, a screen reader (a software program that reads aloud text displayed on a computer screen, or displays it in braille) cost more than $1,000; now, the VoiceOver screen reader is available on every iOS device for free, even the Apple watch," he says.

Smartphone apps make it easier for the blind and visually impaired to navigate other difficult situations, like counting paper money and labeling their surroundings.

Affordable smartphone apps have empowered the blind and visually impaired. Now, artificial intelligence (AI) is taking those apps' capabilities to the next level.

AI and machine learning technologies, specifically computer vision, have grown sufficiently robust to improve the lives of the blind and visually impaired.

Computer vision is a machine's ability to recognize objects in an environment. Microsoft is leading the way with its Seeing AI smartphone app, which uses computer vision to process the immediate environment, other peoples' facial expressions, and text, and then tells what is happening around them. Users simply hold up their smartphones, and the intelligent camera app "sees" for them.

Another example of AI at work is the captioning built into tools like Skype (owned by Microsoft) and Google Slides. The technology uses another application of AI and machine learning, namely natural language processing, to understand the text in a chat or on a slide, then read it aloud. This happens in real time, allowing the blind and visually impaired to navigate conversations and presentations with less friction.

While AI has unlocked promising applications, it is still early days, says Bigham, who points out that AI technologies and tools still make plenty of errors, and few people would want to rely on them exclusively to navigate the world. "We are at the 'better than nothing' stage of AI in accessibility," he says. "That's honestly a lot better than the nothing many people had in some situations not too long ago."

New Tech, Same Problems

While smartphone and AI technology have advanced in the last decade, less progress has been made on the accessibility of our current digital infrastructure.

According to the World Wide Web Consortium (W3C), many sites and tools online today have design barriers that make them impossible for the blind and visually impaired to use, because they don't follow accessibility best practices. In the U.S., it's a legal requirement under the Americans with Disabilities Act (ADA) to make sites accessible, but accessibility rules are unevenly enforced.

"Digital accessibility is still incredibly limited," says Carnegie Mellon's Steinfeld. "It is critical that Web and software companies implement and support established accessibility features to ensure full access to their products."

W3C offers a host of guidelines on making websites accessible for the blind and visually impaired. These include notions like providing text alternatives to non-text content so content can be converted into large print or braille. It also includes making all functionality on a website accessible using only a keyboard.

The W3C organization even has a set of Web Content Accessibility Guidelines (WCAG), a series of documents that "explain how to make web content more accessible to people with disabilities," according to the organization's website.

As one example of what this looks like in practice, many sites are not marked up correctly with HTML, so text-to-speech software cannot accurately translate words into audio for the blind and visually impaired. In most cases, these are easy technology fixes, but they often are overlooked or neglected.

"We have known how to make computing technologies accessible to people who are blind for decades," says Bigham. "Yet the biggest problem remaining is to convince people to code their websites and other applications correctly so they are accessible."

Bigham does note that companies such as Apple, Google, and Microsoft are getting much better at handling issues of accessibility. All three now have large groups dedicated to accessibility issues relating to their products and digital experiences.

For instance, Apple uses its VoiceOver capabilities to describe what is happening on your Apple device. The company also offers visual filters to help the colorblind and magnifiers to help the visually impaired. Apple devices can also be navigated entirely through voice commands, if needed, using Voice Command capabilities.

Artificial intelligence is taking the ability of affordable smartphone apps to empower the blind and visually impaired to the next level.

Bigham notes the accessibility groups at these three major tech companies have "varying levels of influence," and companies like Amazon, Facebook, and Twitter are lagging behind when it comes to accessibility.

Holding technology companies and developers accountable is important work. It also perfectly highlights how frustrating it can be to balance getting existing technology providers to follow the rules, while also inventing better technologies to increasingly enable the blind and visually impaired.

"I would personally prefer to be working on exciting new accessibility applications, but I spend a lot of my time arguing with organizations to make their websites accessible," says Bigham. C

Further Reading

Charters, L.
Apps now online to aid patients with visual impairment, *Ophthalmology Times*, Jun. 5, 2020, https://bit.ly/2GPWxTC

Henry, S.
Accessibility Fundamentals Overview, World Wide Web Consortium, Feb. 19, 2020, https://www.w3.org/WAI/fundamentals

Mobile Fact Sheet, Pew Research Center, Jun. 12, 2019, https://www.pewresearch.org/internet/fact-sheet/mobile/

Cisco Annual Internet Report (2018-2023) White Paper, *Cisco*, Mar. 9, 2020, https://bit.ly/3djrgVH

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

ACM Member News

CONSIDERING THE SOCIETAL IMPACT OF COMPUTING

"In college, at my school, students had the choice to take either the biology track or the computer science track," recalls Kavita Bala, dean of the Faculty for Computing and Information Science at Cornell University. "If you took the biology track you had to dissect a frog, so I picked computer science."

After that, "I was hooked," she adds, "there was no turning back."

She earned her Bachelor of Technology degree in computer science at the Indian Institute of Technology in Bombay, India, and her master's and Ph.D. degrees, both in computer science, at the Massachusetts Institute of Technology. After obtaining her Ph.D., she joined the faculty at Cornell, and has remained there since.

Bala's research interest is in the area of computer graphics and computer vision, with a focus on research in visual recognition, physically based rendering, and material modeling and perception. Her research on scalable rendering technology has been adopted in industrial products for virtual design and prototyping.

In 2015, Bala co-founded the start-up GrokStyle, which created an app that automatically identifies furniture and home decor from just about any picture or angle. Ikea became a client of the company, which was acquired by Facebook in 2019.

As dean, Bala feels it is necessary to consider the societal impact of computing. "We will have to grapple with data science, and information and computer technology, within a societal context, whether it is fairness or access," she said. "These things are going to be more important within the next decade."

—John Delaney

DOI:10.1145/3430368

Cansu Canca

▸ **Susan J. Winter,** Column Editor

Computing Ethics
Operationalizing AI Ethics Principles

A better ethics analysis guide for developers.

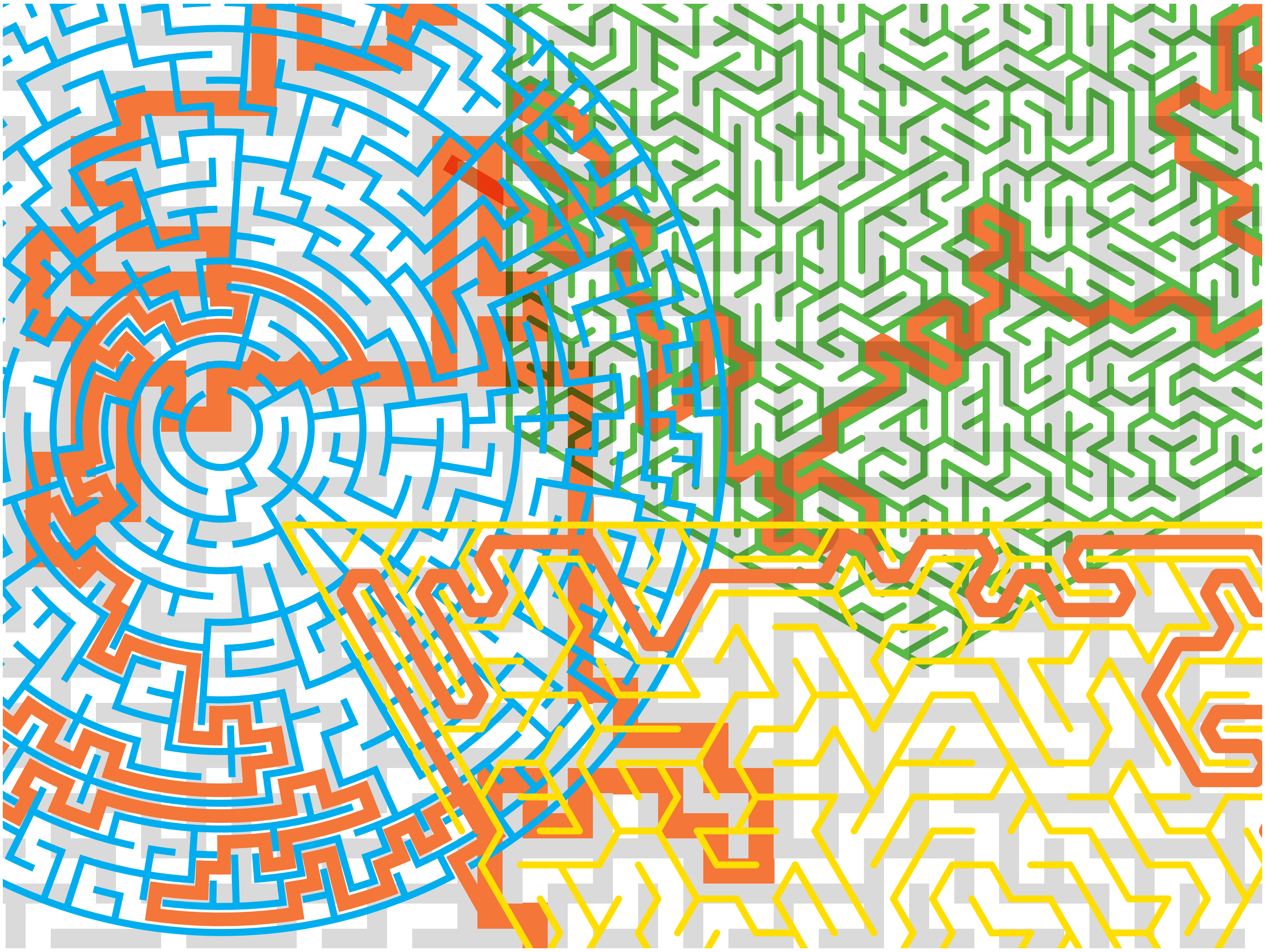
ARTIFICIAL INTELLIGENCE (AI) has become a part of our everyday lives from healthcare to law enforcement. AI-related ethical challenges have grown apace ranging from algorithmic bias and data privacy to transparency and accountability. As a direct reaction to these growing ethical concerns, organizations have been publishing their AI principles for ethical practice (over 100 sets and increasing). However, the multiplication of these mostly vaguely formulated principles has not proven to be helpful in guiding practice. Only by operationalizing AI principles for ethical practice can we help computer scientists, developers, and designers to spot and think through ethical issues and recognize when a complex ethical issue requires in-depth expert analysis. These operationalized AI principles for ethical practice will also help organizations confront unavoidable value trade-offs and consciously set their priorities. At the outset, it should be recognized that by their nature, AI ethics principles—as any principle-based framework—are not complete systems for ethical decision-making and not suitable for solving complex ethical problems. But once operationalized, they provide a valuable tool for detecting, conceptualizing, and devising solutions for ethical issues.

With the aim of operationalizing AI principles and guiding ethical practice, in February 2020, at the AI Ethics Lab we created the *Dynamics of AI Principles*,[a] an interactive toolbox with features to (1) sort, locate, and visualize sets of AI principles demonstrating their chronological, regional, and organizational development; (2) compare key points of different sets of principles; (3) show distribution of core principles; and (4) systematize the relation between principles.[b] By collecting, sorting, and comparing different sets of AI principles, we discovered a barrier for operationalization: many of the sets of AI principles mix together *core* and *instrumental* principles without regard for how they relate to each other.

These philosophical theories answer the central question of applied ethics: What is the right/good action or policy

[a] AI Ethics Lab, *Dynamics of AI Principles*, published in February 2020, https://bit.ly/3k7VgpN

[b] Similar efforts have been done before and we have cross-checked our list with these other works. See Fjeld et al., "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI," *Berkman Klein Center Research Publication 1* (2020), https://bit.ly/2T2sivG; Jobin, Ienca, and Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence 1* (2019), 389–399, https://go.nature.com/3jcvtMc; and Zeng, Lu, Huangfu, "Linking Artificial Intelligence Principles," *AAAI Workshop on Artificial Intelligence Safety* (2019), https://bit.ly/3j6fjUj. These works look at the frequency of principles rather than categorizing them conceptually.



IMAGERY FROM SHUTTERSTOCK.COM

In any given set of AI principles, one finds a wide range of concepts like privacy, transparency, fairness, and autonomy. Such a list mixes core principles that have intrinsic values with instrumental principles whose function is to protect these intrinsic values.[c] Human autonomy, for example, is an intrinsic value; it is valuable for its own sake. Consent, privacy, and transparency, on the other hand, are instrumental: we value them to the extent they protect autonomy and other intrinsic values. Understanding these categories and their relation to each other is the key to operationalizing AI principles that can inform both developers and organizations.

Core versus Instrumental Principles

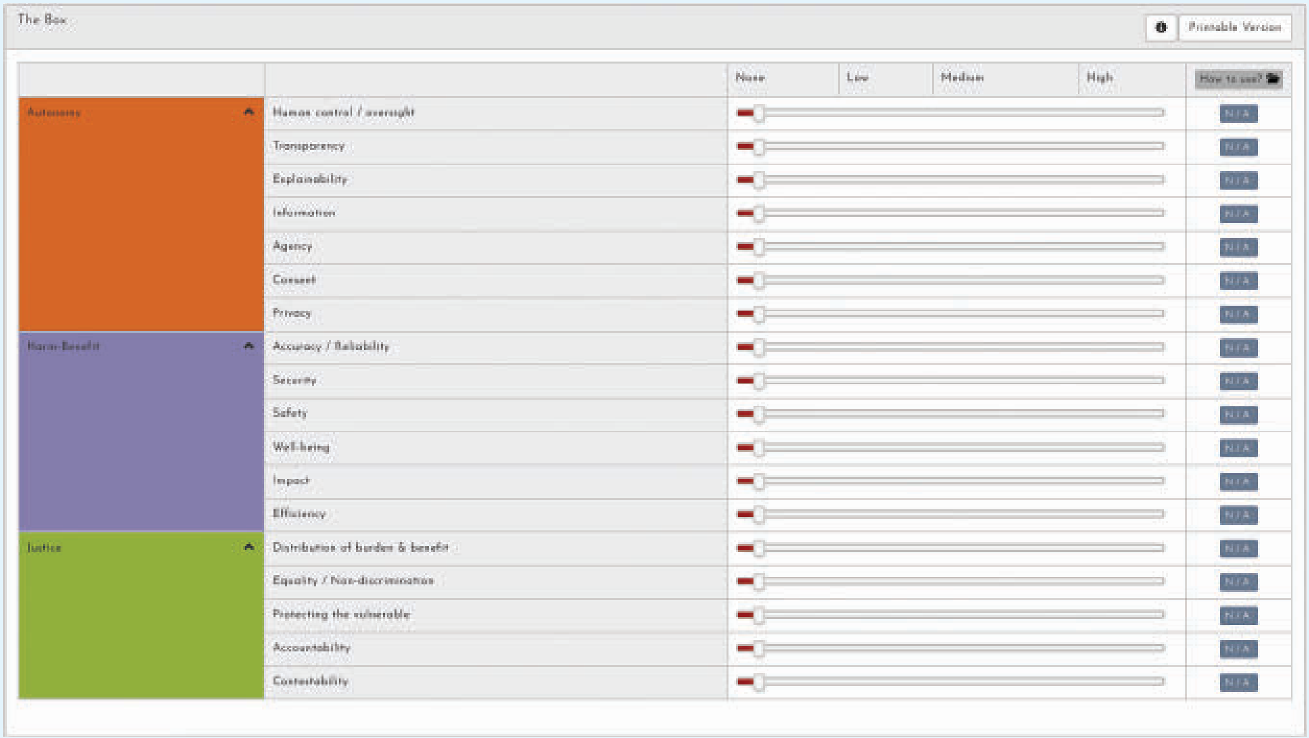
The most widely utilized set of core principles in applied ethics is: respect for autonomy, beneficence (avoiding harm and doing good), and justice.[d] These philosophical principles prescribe an appropriate attitude toward certain values: respect autonomy, do good, ensure justice. They are *core* principles because they invoke those values that theories in moral and political philosophy argue to be intrinsically valuable, meaning their value is not derived from something else. These theories answer the central question of applied ethics: "What is the right/good action or policy to choose?" By encapsulating these theories' intrinsic values in an attitude-setting format, core principles help us immediately recognize if we are facing an ethical challenge: Any action that disrespects autonomy, inflicts harm, or discriminates is ethically problematic, if not straightforward unethical.

When we categorized all of the published AI principles into these three core principles, we found a surprisingly balanced and consistent picture. Across industries and regions, similar weight is given to each of these principles and neither one really outweighs the others: As of

[c] The only document that conceptually categorizes principles is the paper by AI4People, which recites the four "core" principles (beneficence, non-maleficence, autonomy, and justice) and adds the principle of explicability as an "enabling" principle. Floridi et al., "AI-4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, *Minds & Machines 28*, (2018), 689–707; https://bit.ly/343poNt

[d] In 1978, U.S. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research published the *Belmont Report*; https://bit.ly/3dwyOV5. The *Report* laid out three core ethical principles for human subject research: respect for persons, beneficence (which divides further into two general rules of "do not harm" and "maximize possible benefits and minimize possible harms"), and justice. In 1979, philosophers Tom Beauchamp (who was at the Commission co-authoring the *Report*) and James Childress published the canonical book *Principles of Biomedical Ethics*, where they identified four prima facie ethical principles: respect for autonomy, beneficence, non-maleficence, and justice (see Beauchamp and Childress, *Principles of Biomedical Ethics*, 8th edition, Oxford University Press, 2019). Principles in the book and in the report overlap in their content and form what is now often called the "traditional bioethics principles."

The Box.



The Box

🛈 Printable Version

		None	Low	Medium	High	How to use?
Autonomy	Human control / oversight					N/A
	Transparency					N/A
	Explainability					N/A
	Information					N/A
	Agency					N/A
	Consent					N/A
	Privacy					N/A
Harm-Benefit	Accuracy / Reliability					N/A
	Security					N/A
	Safety					N/A
	Well-being					N/A
	Impact					N/A
	Efficiency					N/A
Justice	Distribution of burden & benefit					N/A
	Equality / Non-discrimination					N/A
	Protecting the vulnerable					N/A
	Accountability					N/A
	Contestability					N/A

October 2020, approximately 25%–30% of all principles are autonomy-focused, about another 32%–36% are focused on avoidance of harm and maximizing benefits, and approximately 36%–42% of principles are justice-focused.[e]

In contrast to core principles encapsulating intrinsic values, *instrumental* principles build on concepts whose values are derived from their instrumental effect in protecting and promoting intrinsic values. Take transparency for example. We do not think that something is valuable in and of itself solely because it is transparent. Rather transparency is valuable because it allows us to understand, engage with, and audit the systems that affect us. In other words, transparency is instrumental to uphold intrinsic values of human autonomy and justice. Similarly, accountability in itself is not an end but rather it is a means to safeguard justice by assigning responsibility and to avoid harm through deterrence.

Operationalizing the AI Principles to Guide AI Practice

To guide AI practice, it is important to distinguish core and instrumental principles because instrumental principles are interchangeable. Individuals and organizations can weigh instrumental principles to determine which to prioritize and how to use them to best achieve the core principles.

For example, the instrumental principle of explainability may be required or optional for ethical AI depending on the situation. Explainability is about understanding a system's technical process and how it reaches its outcome. It is an instrumental principle that can uphold human autonomy by allowing individuals to interact meaningfully with the system. It can also help minimize harm and safeguard justice by making it easier to detect errors and biases. Explainability is crucial for a risk analysis AI system that helps judges set bail and parole in the criminal justice system. Here, explainability enables judges and defendants to engage autonomously with the system and better monitor its fairness. In contrast, explainability is not necessary for an AI system that optimizes energy use in cars. Drivers need outcome- and safety-related information to exercise their autonomy, and developers need rigorous testing for safety and accuracy to minimize harm. Explainability in this case should not be prioritized especially if it would compromise accuracy and safety.

The general point is there is no deep ethical dilemma when an instrumental principle is not suitable for a given case. Once we correctly categorize principles as core and instrumental, we can turn many vague AI principles into an operational checklist to guide practice. We created a simplified interactive checklist called *The Box*[f] for computer scientists, developers, and designers to use for basic ethics analyses (see the figure here). The Box helps them determine the relevant ethical concerns and weigh applicable instrumental principles to determine how to best satisfy core principles by substituting or supporting one instrumental principle with another.

The Box also serves as a tool for computer scientists, developers, and designers to recognize when an ethical conflict

[e] The principle of justice holds the largest pie (approximately 40%). This is not surprising because the justice principle is much more general than the other two principles referring as it does to various theories of justice and their main fairness claims such as equal treatment, equal opportunity, and protection of the worst off.

[f] AI Ethics Lab. "Tool: The Box." Toolbox: Dynamics of AI Principles (June 2020); https://bit.ly/2HKKqZ5

is between core principles. Since each core principle is intrinsically valuable, we cannot simply ignore one for the sake of another. When core principles point in opposite directions, we face a real ethical dilemma. In these cases, an ethical issue is complex and cannot be easily resolved solely based on guidance from principles; ethics expert should be brought in to apply ethical theories. An example could be developing an AI system for diagnosing a dangerous and highly contagious disease. To minimize harm and suffering, scientists need to quickly create a highly accurate system. To train the algorithm they need large amounts of personal data that cannot be completely anonymized. Asking for proper consent would delay the project causing more infections and more suffering. However, circumventing consent and disregarding privacy would be a violation of individual autonomy. An AI ethics framework that relies solely on principles is unable to solve this ethical dilemma suggesting that this complex problem should involve ethics experts who can conduct an in-depth analysis and apply ethical theories. Principle-based frameworks only provide a list of considerations rather than a complete and coherent decision-making tool.[g] Ethical theories, on the other hand, provide comprehensive guidance for decision making and ethics experts can utilize these detailed theories to analyze complex issues in-depth.

When properly operationalized, AI principles provide a helpful start for an ethics analysis and they can guide developers and organizations through many ethical questions, even though they are not sufficient for complex ethical problems. We need to understand AI principles for what they are: A list of fundamentally and instrumentally important ethical considerations but *not* a complete system for complex ethical decision making. Categorizing and using AI principles ensures we do not overlook a crucial ethical concern. By revealing when a case presents a conflict between core principles rather than instrumental ones, principles can also help us recognize when we face a complex case and need a full-scale ethics analysis. As we operationalize AI principles, we need to utilize their strengths and recognize their limitations, acknowledging that AI principles are only a first step for development and deployment of ethical AI.

Principle-based frameworks only provide a list of considerations rather than a complete and coherent decision-making tool.

Setting an Organization's AI Principles

Lastly, let us go back to the proliferation of AI principles. What is the point of company-specific sets of principles if the content is largely the same, as we have seen? If well done, the point is that organizations with their own customized set of AI principles can determine how to weigh competing principles and which intrinsic value to prioritize when core principles (and theories) conflict. When no single argument emerges as the strongest even after a full-scale ethics analysis, developers and other organizational decision-makers must choose between equally ethically permissible options. This is when the organization's stance about core principles shows through. When push comes to shove, does this organization prioritize individual autonomy or minimization of harm? An organization's AI principles can guide this decision and be useful both for computer scientists and ethics experts when they are deciding a hard case. To help them clarify their own values, we invite organizations to use our toolbox to compare their efforts to others, systematize their AI principles, and engage in in-depth ethics analysis with experts to determine their own priorities for coherent and consistent ethical decision making.

[g] An excellent paper on this is Clouser and Gert, "A Critique of Principlism," *Journal of Medicine and Philosophy 15*, 2 (1990), 219–236; https://bit.ly/342fITB

Cansu Canca (cansu@aiethicslab.com) is a philosopher and the Founder and Director of the AI Ethics Lab in Boston, MA, USA.

DOI:10.1145/3430375

W.R. Adrion, S. Dunton, B. Ericson, R. Fall, C. Fletcher, and M. Guzdial

▸ Richard Ladner, Column Editor

Broadening Participation
U.S. States Must Broaden Participation While Expanding Access to Computer Science Education

Incorporating equity and inclusion in the effort toward access for everyone.

MAKING SWEEPING CHANGES to education in the U.S. is difficult because of its highly decentralized primary and secondary school system. Each of the U.S. states and territories makes its own decisions about how education is structured. Some of those states push the decision-making to districts and even individual schools. Reforms, such as providing high-quality computer science (CS) education to *all* students, require states to engage every school district, if not every school. In order to broaden participation in computing, rather than exacerbate existing inequities as we expand K–12 CS education to more U.S. schools, explicit attention needs to be placed on how equity is addressed and measured in policy, practice, and professional development.

Many states are making progress on CS education. As of October 2020, 18 states have started or completed statewide plans; 37 states have defined CS standards; 40 states plus the District of Columbia have teacher certification for CS.[2] The Expanding Computing Education Pathways (ECEP) Alliance[a] is one of eight Broadening Participation in Computing Alliances[b] funded by the National Science Foundation. Begun in 2012 (and described a previous *Communications* Education column, "Broadening Access to Computing Education State by State," in February 2016), ECEP focuses on state-level educational systems and now works with 22 states and the territory of Puerto Rico to ensure the goal of *broadening participation in computing* (*BPC*) is a priority.[1] Supported by additional NSF funding in 2018, the ECEP 2.0 leadership team continues to build on the early work of ECEP while scaling the model of state-level BPC work beyond the ECEP states.

ECEP member states are making great strides toward increasing the number and diversity of students who have access to high-quality computing education. But often, there is more pressure to increase the *number* of students who have access than to

[a] See https://bit.ly/31APN3V

[b] See https://bit.ly/37h074r

PHOTO BY NEONBRAND ON UNSPLASH

ensure the *diversity* of those students. It is easier to increase access to computing education in well resourced schools, which tend to be more homogenous. Increasing access, participation, and experience to students who have been historically underrepresented in computing[c] requires broad-based teams who are committed to making change through data-driven strategic efforts.[5]

Why is it difficult to develop and maintain goals for BPC in state education systems? First, such goals are hard to define. How can progress be measured? What data is informing the development of goals? How do we know that changes, be they state education policies or classroom practices, result in engaging a more diverse set of students in CS? Also some goals, while they seem ambitious, are not enough to broaden participation. Ensuring there is a CS teacher in every school is not enough. Many schools are highly diverse, but the computer science classes tend to be mostly White or Asian and male. If the CS teacher offers one small elective class, most students still do not have access. Making computer science classes a requirement is not enough. Many states are enacting a requirement that CS classes be made available in every school, but are not providing funding for teacher professional development nor are intentionally defining what content makes a course a CS course. Schools can often meet this requirement by simply providing access to online classes. However, historically marginalized students face numerous barriers to success in online courses, so the result is still too few students getting access.[3]

The historic inequities embedded in our social systems make the challenges in reaching "CS for All" enormous, but not insurmountable. COVID-19 and the onslaught of virtual educational, professional, and community spaces have brought new awareness to our dependence on technologies and the computer science that is inherent in every aspect of our lives. The pandemic also has magnified the inequities faced by families with limited technology access or literacy and limited skills or time to support their students' online learning. In many communities, parents or guardians do not see technology careers as achievable pathways for their children.

Computer science may be an aspiration, but it is not a priority for many school administrators and principals.[4] School leaders find themselves with a limited understanding of *what* CS is and *why* all students need to learn it. Pressures to address traditional, tested curricula, and a lack of resources and teachers to meet the demand for a new discipline further compound the challenges. The pressures placed on schools, school leaders, teachers and students by standardized testing of reading, writing, mathematics, and science, particularly in schools where large numbers of students struggle with such tests, can result in lower priority for computer science curricula. The "time in the day" for traditional subjects can push CS off as an elective. The pandemic has increased online learning, limiting hours of interactive learning and exacerbating the availability of time CS curricula. CS teachers are hard to find and retain. Too few university programs are training potential computer science teachers. Professional development for in-service teachers is limited. Teachers are often shifted from teaching CS to perceived higher priority subjects such as math or science. The outcome is too often restricted opportunities for students who have had historically limited access to learning the computing skills they need to succeed in any career, much less having been encouraged into computer science pathways. Computer science education must not be seen as yet another burden on schools and teachers. It is beyond time that we flip the educational script and elevate "CS for All" as a priority and see its diversity, equity, and inclusion goals as an opportunity to equip all graduates with the skills they need to succeed in the 21st century.

Several ECEP states are rising to this challenge by putting broadening participation in computing goals at the center of their advocacy and policy efforts. They are creating plans for

[c] See "Women, minorities (African Americans/Blacks, Hispanic Americans, American Indians, Alaska Natives, Native Hawaiians, Native Pacific Islanders, and persons from economically disadvantaged backgrounds), and persons with disabilities" from https://bit.ly/2Hl0uzZ

growing CS education and broadening participation at the same time. A primary focus for ECEP 2.0 is to support states to develop ways to disaggregate and track data that measure BPC goals. Because computing is a new subject for many states, it has been difficult to define computing education and even more difficult to measure the demographics of students engaged in computing. Due to concerns about confidentiality and masking identifiable information when student numbers are small, it has been challenging to disaggregate and report on ethnicity, gender, and disability. As a collective impact network, ECEP is helping state leaders develop consistent, replicable strategies for analyzing state data through the lens of BPC. Two examples, from Rhode Island and Nevada, where state goals and plans explicitly aim to increase the number *and* diversity of students getting access to high-quality CS education, are described here.

Rhode Island

In 2016, Rhode Island Governor Raimondo launched Computer Science for Rhode Island (CS4RI)[d] with the aim of having computer science taught in every public school. CS4RI brought together state government and its education department, K–12 public education, universities, industry and nonprofits. The state provided more than $200,000 per year for the past three years to support staff, events, and teacher professional development. CS4RI's team selected providers of CS educational content, curriculum, and teacher professional development that would support RI's initiative to bring computer science to all students. One of the guiding principles when the state developed CS standards stated: "All students regardless of age, race, ethnicity, gender, socioeconomic status, special needs, English proficiency, or any other demographic will have the opportunity to participate in computer science. The content and practices of the standards will be accessible to all."[e]

Rhode Island's strategy for broadening participation in computing relies on four key actions: identifying a member of the Core Team to serve as the diversity coordinator; requiring that all state-approved CS content providers meet a rubric for culturally responsive and accessible curricula and professional development; offering an online diversity course for all RI CS teachers; and partnering with research projects to identify the level of engagement of students with disabilities. CS4RI also supports districtwide planning workshops for CS educators and administrators that include inclusion and diversity training.

Rhode Island's efforts are showing results. By the end of 2017, the state had reached its goal of having CS offered in every public elementary, middle, and high school in at least one of three modes: standalone courses; CS lessons integrated into existing courses; or access to CS courses during the school day from another Rhode Island school. By the end of 2018, 46% of high schools were offering AP computer science, a 26% increase from 2015. Female students in AP courses rose from 5% in 2015 to 30% in 2018. And the number of female CS major college graduates in the state rose from 17% to 25% over the same period.

Nevada

Nevada's Board of Education vice president Mark Newburn and the STEM Coalition K–12 CS Task Force leveraged membership in ECEP to create a CS coalition that built a cohesive statewide CS education expansion effort. Over two years, NV leaders contributed to the national K–12 CS Framework and also developed a process for writing CS standards for NV. The broad-based team framed a strategy to broaden participation and built statewide support for CS education. Concurrently, the Nevada legislature identified gaps in K–12 CS education instruction and drafted Senate Bill 200. Passed in June 2017, the bill requires all public high schools, charter schools, and university schools for gifted students to offer a state-approved CS course by 2022. Notably, the legislation[f] states: "These schools must also make efforts to increase enrollment of girls, students with

[d] See https://bit.ly/2T5IYT3

[e] See https://bit.ly/3m3yjof

[f] https://bit.ly/3oQcUkR

disabilities, and underrepresented minorities in the field of computer science as identified by the state board." Senate Bill 200 also mandates that Nevada's half-credit Computer Literacy graduation requirement must now include at least 50% of computer science content, thus guaranteeing all students receive some exposure to computer science concepts.

After the bill was passed, NV ECEP leaders and state department of education staff organized a statewide CS education summit at which stakeholders from K–12 and postsecondary education, government and industry learned strategies to support equitable CS education and wrote strategic plans that focused on broadening participation, which were collected and shared. The state education department also followed up with an online seminar series and training for school counselors.

Nevada allocated $1.4 million in state funding to districts and charter schools for teacher and school leader professional development, curriculum, and equipment. Nevada's goal to broaden participation was integrated into how it distributed state funds. Each applicant had to show how funds would be used to increase the enrollment of females, students with disabilities, and underrepresented minorities. Of the 17 districts in NV, 14 applied for funding, and nine charter schools were funded as well.[6]

Nevada's commitment is showing dramatic results. From 2016 to 2019, the number of students taking computer science doubled and the number of high schools offering computer science more than tripled. From 2017 to 2019, Nevada students taking the Advanced Placement Computer Science Principles exam increased by fivefold. Hispanic students' represented 40% of all students in 2019 (a more than eight times increase over 2017), and female students represented 38% of Nevada's CSP exam-takers, up from 30% in 2017.[g]

Nevada continues to focus on broadening participation as it creates a state strategic plan, pre-service and in-service professional development for K–12 teachers, and guidance documents. Nevada's CS team's work has never been about the number of students reached, it has been intentionally and systematically focused on offering high-quality computing education to all students and reaching those underrepresented in computing.

In the movement toward "CS for All," state K–12 education systems often focus on access, but too often equity and inclusion are afterthoughts in policy and strategy efforts.

Conclusion: Getting Access to Everyone

In the movement toward "CS for All," state K–12 education systems often focus on access, but too often equity and inclusion are afterthoughts in policy and strategy efforts. Even mandates to offer CS in every school may not mean that students with disabilities are fully included, or that every girl participates, or that students in schools that offer online CS courses succeed in those courses. As states enact new policies, we recommend the following actions:

- Include people with expertise in diversity and inclusion on the team developing policies and plans. Broad-based teams bring a wide range of perspectives that are vital to serving a wide range of students.
- Tie resources to goals to broaden participation. Provide funding to all districts, not just the well-off schools that are poised to take advantage of competitive grant funds. Ask districts to show how new resources will increase participation by historically underrepresented students.
- Ensure state-approved CS curriculum and professional development incorporates strategies to broaden participation and enable all students to succeed, no matter the tools or delivery mode.

As members of ECEP, Rhode Island and Nevada have developed and shared their approaches to expanding CS education in ways that focus on broadening participation among students underrepresented in computing fields. The 22 ECEP states and the territory of Puerto Rico are working hard to ensure broadening participation in computing remains a top priority and that the policies and resources target that goal.

[g] See https://bit.ly/2T5IKeF

References

1. Adrion, R. et al. Broadening access to computing education state by state. *Commun. ACM 59*, 2 (Feb. 2016), 32–34.
2. Code.org. State Tracking 9 Policies (Public), (Oct. 2020); https://bit.ly/3dDUq1H
3. Dynarski, S. Online courses are harming the students who need the most help. *New York Times.* (Jan. 19, 2019); https://nyti.ms/3dFnmpX.
4. Herold, B., and Schwartz, S. Principals warm up to computer science, despite obstacles. *Education Week 37*, 27 (Apr. 18, 2018), 24–25; https://bit.ly/31gakKz
5. Ladner, R. and Israel, M. Broadening participation "for all" in "computer science for all." *Commun. ACM 59*, 9 (Sept. 2016), 26–28.
6. Nevada Department of Education. Status Report Senate Bill 200 Computer Science Education. (2019); https://bit.ly/2T7ZIJy

W. Richards (Rick) Adrion (adrion@cs.umass.edu) is professor emeritus at the College of Computer and Information Sciences at the University of Massachusetts Amherst, USA, and was UMass ECEP principal investigator 2012–2019.

Sarah T. Dunton (sdunton@mghpcc.org) is Director of the Expanding Computing Education Pathways (ECEP) Alliance at the Massachusetts Green High Performance Computing Center, Holyoke, USA.

Barbara Ericson (barbarer@umich.edu) is assistant professor of information, School of Information at the University of Michigan, USA, and former co-principal investigator of Georgia Tech ECEP.

Renee Fall (rfall@css.edu) is senior research scholar at the National Center for Computer Science Education at the College of St. Scholastica in Duluth, MN, USA, and former co-principal investigator of UMass ECEP.

Carol Fletcher (cfletcher@tacc.utexas.edu) is director of EPIC at The University of Texas at Austin, USA. She is the current principal investigator of the ECEP Alliance.

Mark Guzdial (mjguz@umich.edu) is professor of electrical engineering and computer science in the College of Engineering at the University of Michigan, USA, and was Georgia Tech ECEP principal investigator 2012–2018.

Thanks to ECEP members from Rhode Island (Carol M. Giuriceo, Rhode Island College and Victor Fay-Wolfe, University of Rhode Island) and Nevada (Cindi L. Chang, Nevada Department of Education, and Mark Newburn, Nevada Board of Education) for their contributions to this column.

This material is based upon work supported by the National Science Foundation (NSF) under Grant Nos. NSF-CNS-1822011, NSF-CNS-1228355, and NSF-CNS-1228352. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Expanding Computing Education Pathways (ECEP) is part of the National Science Foundation's Broadening Participation in Computing Alliances (NSF BPC-A).

DOI:10.1145/3430370

Peter J. Denning

The Profession of IT
Navigating in Real-Time Environments

An interview with Jim Selman.

JIM SELMAN HAS been a professional leadership coach for over 30 years. He frequently encounters executives and team leaders who are dumbfounded because the world is changing so rapidly and sometimes chaotically that their best laid plans are useless and ineffective. Many computing professionals have a similar experience today after the COVID-19 avalanche swept through. In his recent book, *Living in a Real-Time World*, he summarized his conclusions about what professional leaders should learn in order to be effective in this environment. I explored this issue with him.

—Peter J. Denning

DENNING: Your book *Living in a Real-Time World, 6 Capabilities to Prepare Us for an Unimaginable Future* seems particularly relevant given what's happening around the world. What do you mean by 'real-time world'?

SELMAN: I started my career in the 1960s in IT as a programmer and systems analyst. At the beginning, computers were 100% information processing machines. As processor speeds got faster and faster, the gap between inputs and outputs got shorter and shorter to the point of being imperceptible—computers transformed from informing machines to performing machines. The idea of "real-time" computing and capabilities like process-control came into existence. Real-time computing means computing that responds rapidly and effectively to inputs as they appear, without knowing when or if they will occur.

I'm using the term in the same way. The technology-charged world is changing so rapidly that many of our plans and expectations are dashed by surprises. The future we imagined and planned for never appears. This is why I call the future unimaginable. This is immensely frustrating to many people. Today, in my opinion, we need to be more like the real-time computers, responding to what actually appears rather than what our best laid plans expect to appear. We need to transform our worldview, our practices, and our skills to successfully navigate a reality that is increasingly unpredictable and beyond our control.

What do you think is driving all this chaos and unpredictability?

I don’t know why the world is the way it is or what causes anything really given the overwhelming complexity we’re experiencing. I suspect that computers, networks, and the whole of technological progress is a big part of what has accelerated the pace of change beyond our comprehension. The massive increases in efficiency and knowledge over the past few decades have been amazing and beneficial. Yet, as you pointed out in your last piece on the current “avalanche” of change, the world economy is subject to “avalanches”—disruptive changes that sweep away jobs, identities, wealth, and opportunities. An avalanche leaves people feeling lost, confused, left behind, or afraid. While many avalanches are small, affecting only a limited segment of the economy, COVID-19 is an avalanche that no one could escape. It triggered other avalanches such as the collapse of some industry sectors (for example, air transport), oil prices, international trade, and some higher education systems. Like it or not, this avalanche calls us all to navigate unimagined environments.

The ‘Four Horsemen” of Silicon Valley (Amazon, Google, Facebook, and Apple) are leading the charge into a future that is unimaginable. But the current pandemic, threats from climate change, and unprecedented levels of unemployment make us realize this *is* the unimaginable future—a perfect storm of disruptive change. Why we got here isn’t very relevant. The only question is what to do now—how do we navigate, make choices, plan and invest in the future when we don’t trust our predictions and we have little control over changes that we cannot comprehend?

I am fond of Star Trek as a metaphor for this situation .... we’re “going where no one has gone before” and we have no certainty of our destination or maps to guide us. I suggest that our traditional notions of leadership are of necessity transforming from “Leader” to “Navigator.” Navigators don’t know any more than anyone else, but they keep us centered in where we are and where we’ve come from. With that we can move and shape the new world.

What are the traditional ways we have coped with disruptive change? Why aren’t they working anymore?

Historically, we’ve dealt with change pretty much the way we deal with everything else. We learn from the past and we gather techniques and recipes. We apply all that to the current situation in the interest of controlling future outcomes or solving specific problems. This mindset doesn’t work in a real-time world. When you make decisions and commitments and allocate resources based on an assumption that the future will be much like the past, you only guarantee dissatisfaction and frustration when that future does not appear. This simply creates a counterproductive vicious cycle that worsens the situation and is often self-destructive.

> **“Computer professionals in the future need to be philosophers, leaders, and navigators in addition to being technical experts.”**

What do we need to learn in order to regain our effectiveness in this kind of environment?

In my view, human beings already are fully equipped and have all the capabilities to become successful navigators to succeed in a world of permanent uncertainty. For example, everyone has an ability to read even if they are initially illiterate. What are the capabilities inherent in all of us that allow us to be effective in a real-time world? In my book, I mention six of these inherent capabilities:

▸ **Accepting** or what I call the Art of Surrender. This means accepting the world as it shows up. Acceptance cultivates our capacity to choose. When you are resisting or reacting you are not choosing. When you have discovered the impossibility of winning the game you are playing you can choose to surrender. You can’t play a new game until you give up the game you are playing. Surrender is a choice. It isn’t succumbing or being defeated.

▸ **Being** or what I call the Art of Context. Computer professionals know better than most that computers operating without context are limited to purely mechanical operations. In the human domain, context is all-important. It gives us meaning and purpose and is the key to have freedom and power. In a real-time world we need to be continuously aware of the fact that we have a choice about our way of being—the embodied way we relate to our context. When you shift your way of being, you shift the context in which you observe and that will always open new choices and possibilities that you had previously been unable to perceive.

▸ **Listening** or what I call the Art of Mastering Moods. I am not talking about hearing sounds but something much deeper—our background of awareness and interpretation of what other people are saying. We call this background our mood. Our mood orients us to interpret the world in particular ways. Our mood includes our stereotypes, prejudices, and “mental models.” Our moods, our thinking, and ultimately our behaviors are inseparable from our listening. If we want to be successful navigators, we must become aware of our moods and shift them to ones that accept the world as it is and give us choice about what we do next.

▸ **Communicating** or what I call the Art of Relating. Everyone knows the importance of communication and relationship. Not everyone appreciates that communication and relationship are two sides of the same coin and are learnable skills. In a real-time world mastering how we relate to others, how we relate to our circumstances, and even how we relate to ourselves becomes essential if we’re to stay centered in the flow of life and be free to continually choose our actions.

▸ **Appropriating** or what I call the Art of Situational Learning. Most things are changing faster than we can comprehend and what we’re learning is often obsolete before we learn it fully. This is especially obvious in the technical fields. In the distant past, human beings viewed learning as a team effort and the idea of an individual being the container of knowledge made no

sense. Today there is so much information available to us that it still makes no sense that a single person can contain all knowledge. Today we can rebuild practices for spreading the learning across multiple people and then bringing it all together through collaboration and focus on an objective.

▸ **Caring** or what I call the Art of Love. One aspect of the emerging reality is the recovery of our appreciation for each other and the planet. Every age is characterized by core values and ideals. For most of the last few centuries the focus has been on production and control. In our real-time world, there is a shift from these to what I believe is more fundamental and relevant—care. Nothing matters without care.

These responses may strike some people as "stuff we all learned in kindergarten." Or perhaps as "soft" and lacking in rigor. You are saying we really didn't learn this in kindergarten or any other part of our schooling. Why do you say this? Why is your framework rigorous?

At the end of the day our "reality" is a function of action—our future will be shaped by our actions. One of the premises in my work is if something isn't observable, it isn't actionable. This is why I prefer a more phenomenological approach in which we see we swim in an ocean of language, not a pool of words. We inhabit networks of conversations over which we have little visibility or control. The skill I want to cultivate looks closely at language to see where action is created in conversations. Action is created by commitments. We can be rigorous observers of the commitments we make in our conversations, and whether our actions line up with our commitments.

What is your experience? Are project leaders receptive to the idea of real-time world?

If you believe a new post-avalanche reality is emerging, you must confront many assumptions you've taken for granted but no longer relevant or workable. Even our common sense breaks down. Are people receptive to this interpretation? Often not initially. But most people know our current way of approaching things isn't working so well and something is profoundly missing. They are actually open to the possibility there might be a better way of dealing with things. I don't ask people to believe anything, or agree with anything or even understand everything. I ask only they consider having a serious conversation about what they are already dealing with and be open to testing or trying some of the practices I suggest. It doesn't take long for them to engage and realize they are learning a kind of new language—a language of leadership. It is as if they learned a higher level programming language, which opens choices and possibilities not previously available.

How would this help computing professionals?

It's pretty obvious that computers are expanding exponentially. Look at all the speculation around AI, robots, and virtual realities. Computers may be the most critical element driving us into a real-time world. Historically, information systems and processes are material processes that move signals representing symbols in time and space, in a cause-effect world. The emerging world is different. The big question is where human beings fit into the technological picture. I believe the answer is in the synthesis between the physical and social sciences. Computer professionals in the future, I believe, need to be philosophers, leaders, and navigators in addition to being technical experts. They cannot be relegated into narrow technical niches. To me, this is the future essence of professionalism. C

Suggested Readings

1. Flores, F. and Winograd, T. *Understanding Computers and Cognition.* Addison-Wesley, 1987.
2. Selman, J. *Living in a Real-Time World.* Independently published, 2019; available at amazon.com.
3. Selman, J. Leadership. CreateSpace independent publishing platform, 2016; available at amazon.com.

Jim Selman (jimselman@paracomm.com) is a seminal leader in the theory and practice of business coaching. He contributed several new concepts and techniques to the field of management, notably organizational transformation, coaching, the Merlin method for designing the future, breakthroughs, and breakdowns. He developed new approaches for leaders to producing broad "paradigm shifts." He is a former partner in the firm of Touche Ross (Deloitte Touche), co-founder and CEO of a consulting network, Transformational Technologies, and founder and CEO of Paracomm Partners, Petaluma, CA, USA.

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for information innovation at the Naval Postgraduate School in Monterey, CA, is Editor of ACM Ubiquity, and is a past president of ACM. The author's views expressed here are not necessarily those of his employer or the U.S. federal government.

DOI:10.1145/3430378

George V. Neville-Neil

Article development led by acmqueue
queue.acm.org

Kode Vicious
Removing Kode

Dead functions and dead features.

Dear KV,
Your columns often discuss issues that come up when *adding* new code, but when is the right time to *remove* a piece of code? I have been working on an enormous legacy code base and most of the questions that come up in our stand-up meetings concern which modules we think we can remove from the system.

Ready for Removal

Dear Ready,
I am surprised and amused at the narrowness of your question. Most software developers, when confronted with a legacy code base, simply want to throw it out and start again. I have to say this happens to me at least once a week, just looking around at the astounding mountains of legacy that pass as software.

Removing dead code from systems is one of KV's favorite coding pastimes because there is nothing quite like that feeling you get when you get rid of something you know was not being used. Code removal is like cleaning house, only sometimes you clean house with a flamethrower, which, honestly, is very satisfying. Since you are using a version-control system (you had better be using a VCS!), it is very easy to remove code without worry. If you ever need the code you removed, you can retrieve it from the VCS at will.

As to when you should be removing code, there are several answers. The first, of course, is that if you have truly dead code that can no longer be reached from anywhere in the system and that is not simply conditionally compiled-out test code (see my February 2017 column, "The Unholy Trinity of Software Development") then that code should be removed immediately. A good compiler or other tool will tell you when you have dead code, which should make the job fairly straightforward.

Complications arise when you have code that is infrequently used but has been in the system for a long time. At this point, you have to do a bit more thinking about your code and whether or not the code you are looking at is—for want of a better term—*nearly dead*. Any system that survives for a number of years will tend to grow functionality that is used—sometimes briefly—and then discarded or ignored. The logic for not removing dead features is usually stated as, "If it ain't broke ... " which is actually quite foolish. If your software is so fragile that removing a feature breaks the whole system, you have much bigger problems than needing a bit of code cleanup, and that means the system as a whole probably needs rototilling if not an outright rewrite.

A dead *feature* is different from a dead *function*. Dead features are either completely unused or used by only a minority of users. The risk with dead or nearly dead features is that they leave a larger attack surface in your code. Once upon a time, we also cared that dead features left code in the executable that bloated the system, but only those of us working in the embedded-systems area seem still to care about that. For modern server-based systems, it is the risk of the dead feature giving an attacker a place to infiltrate your system that is probably preeminent in your mind, or at least it should be.

Sometimes a dead feature will have a noisy minority of users or an internal champion who has tended that feature for many years. At this point, the removal becomes a political (that is, human) problem. There are several ways to solve political problems at work, but some of those will get you 10 to 20 years in jail if you carry them out. If the noisy minority is very small, it is possible that feature could be broken out into its own separate program so that the code is no longer part of the larger whole. If the problem is developers, well, it is time to give them a new challenge, far away from their pet features. Once you have developers with pet features, you have a very different sort of management problem, one I will not address here.

You will notice I did not give you a definite timeline in answer to your question, and that is because there really isn't one other than removing truly dead code as soon as it is dead. For feature removal, that really depends on how you deal with your users and developers. A good rule of thumb, though, is to remove features only at a major release so as to limit the level of surprise.

KV

Related articles on queue.acm.org

Code Hoarding
Kode Vicious
https://queue.acm.org/detail.cfm?id=2897034

Surviving Software Dependencies
Russ Cox
https://queue.acm.org/detail.cfm?id=3344149

Writing a Test Plan
Kode Vicious
https://queue.acm.org/detail.cfm?id=3294770

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

DOI:10.1145/3431480

Margaret O'Mara

Viewpoint
Silicon Politics

Tracing the widening path between Silicon Valley and Washington, D.C.

Tech and politics don't mix. That has been the story Silicon Valley leaders have broadcast to the world since the region first sprang into the forefront of public consciousness as the land of silicon chips, personal computers, and video games. It is an attitude in keeping with the celebration of rugged individualism and disdain for centralized political power that has been part of American political culture since the nation's founding, ideas that gained additional allure amid the stagflating malaise of the post-Vietnam, post-Watergate 1970s. In the Reagan Revolution year of 1980, the sole election-year commentary in the microelectronics-industry newsletter *InfoWorld* was a cartoon tucked into a bottom corner of the editorial page. "I was going to keep track of all the candidates' significant statements," one man sighed to another as they stood in front of a computer terminal, "but there's no way to process an empty disk."

Four years later, Steve Jobs declared, without embarrassment, that he had never voted in his life. 1990s-era moguls including Bill Gates and Jeff Bezos ducked questions about their voting preferences for years. In the early 2000s the loudest and most unapologetically political voices coming out of Silicon Valley were libertarians such as PayPal cofounder and Facebook board member Peter Thiel. Politicians of both parties long courted Silicon Valley's affections, but for many in tech, politics was something to be publicly ignored, if not actively disdained.

U.S. President Donald J. Trump meets with Facebook CEO Mark Zuckerberg at the White House in Washington. D.C., in September 2019.

OFFICIAL WHITE HOUSE PHOTO BY JOIYCE N. BOGHOSIAN

Six years ago, as I embarked on the research process for my recently published history of Silicon Valley, *The Code*, I told a longtime tech-industry veteran that part of my goal was to show how tightly the industry's rise and evolution intersected with modern American political history. "Can I tell you something?" my interviewee interjected, not unkindly. "If you write about tech and politics, that book is going to end up on the remainder table. Because there's no story there."

My proposition is not such a tough sell in the America of 2020. Before the COVID-19 pandemic ground normal life to a halt, tech CEOs ricocheted between White House photo-ops and scorching presidential tweets. Many in the tech rank-and-file donated to progressive Democrats like Bernie Sanders and Elizabeth Warren while members of tech's C-suites threw high-dollar fundraisers for Democratic centrists and the Republican incumbent alike. The size and scale of tech's biggest platforms and products triggered sharp public criticism and fresh lawmaker scrutiny. Tech regulation was the rare issue on which the two parties seemed to agree.

This is not as much of a turnabout as it seems. Silicon Valley's mythology of independence to the contrary, politics and government are absolutely central to its story. This was the case in other industrialized nations as well, but only in the U.S. have tech entrepreneurs and many of their allies embraced such a thoroughly market-driven understanding of their accomplishments. American technologists' collective amnesia has served a political purpose of its own, instilling a sense of entrepreneurial agency amid intensive government mobilization and market intervention, and also building a political case for, among other things, self-regulation of online platforms. To understand how and why to grapple with today's tech policy landscape, we must reckon with this history and the institutional reality that the U.S. government built the tech industry, but it did so in a way that helped the industry's leaders believe they did so all on their own.

One point of connection comes with an institutional symbiosis familiar to any technologist who has worked in an NSF-funded laboratory or entered a DARPA robotics competition: the academic-military-industrial complex established in the early Cold War. As historians of science remind us, this purportedly apolitical Big Science was in fact deeply political. Competing with the Soviets provided the rationale for unprecedented public investment in everything from university science and engineering programs to the rockets of the Apollo program.

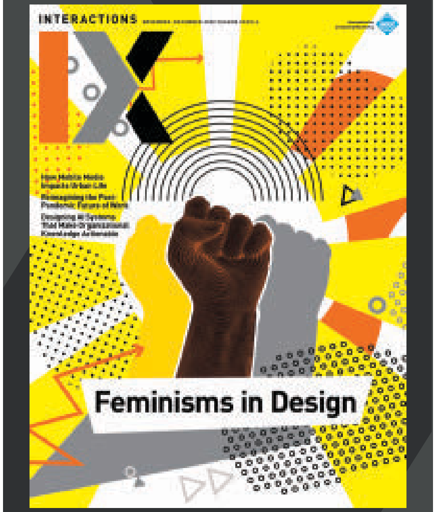
Yet as significant as the money itself was the way in which it flowed: *indirectly*, via grants and contracts to private defense contractors, to research universities, to industrial research labs. This obscured the extent to which government spending was undergirding the enterprise, yet another example of the indirect and privatized process of state-building that has been a hallmark of American federalism since the 19th century. In the early Cold War period, this structure spurred innovation and market competition in advanced electronics industries at a time when commercial demand was nearly nonexistent, creating a foundation for later growth of consumer and business markets.

Take for example the case of the silicon semiconductor industry that gave the Valley its name. Fairchild Semiconductor, the pioneering startup whose many spinoffs include Intel and the legendary venture capital firm Kleiner Perkins, had the lucky break of incorporating in September 1957, only a few weeks before the Soviets launched Sputnik into orbit and sent the American space program into overdrive.

Silicon Valley's mythology of independence to the contrary, politics and government are absolutely central to its story.

Sending satellites into orbit, or astronauts to the Moon, required light, powerful electronics like the semiconductors and integrated circuits developed by Fairchild. The technology was so new that only Fairchild and other specialist electronics firms like it could supply what the space program needed. The spike in demand coincided with budget-minded Defense Department reforms to the prime contracting system that dislodged larger and more well-established firms. The result was that in Fairchild's early years, this granddaddy of all Silicon Valley startups had more than 80% of its book of business come from government contracts. Large purchase orders from the Apollo program for integrated circuits drove down the price and allowed chipmakers like Fairchild to scale up production, enabling a commercial market to emerge by the end of the 1960s.

A similar combination of entrepreneurial energy and government subsidy and encouragement lies behind the origins and evolution of the Silicon Valley-based personal computer industry. Unlike the Cold Warriors who inhabited the early electronics industry, the personal-computer makers hailed largely from a Vietnam generation who wanted nothing to do with the military-industry complex and instead saw computers as tools of liberation from the establishment's rules and gatekeepers. Instead, they crafted an anti-politics that was in itself a political position of its own. Alternatively called "new communalism," "cyberlibertarianism," or "the Californian ideology," this techno-utopian variant of 1960s



countercultural thought was both deeply cynical—governments and traditional authorities were not to be trusted—and earnestly techno-optimistic.[a] Place a computer on every desk, and connect their users in nonhierarchical networks of communication and creation, and you would fix what ails an ailing world.

But the proclamation and dissemination of this ideology obscured the fact that, particularly for those running high-tech enterprises, tight connections to traditional political institutions and political processes remained an essential part of doing business. Valley leaders were active and consistent lobbyists for help from Washington as the industry weathered the economic storms of the 1970s and launched a new set of consumer-facing companies in the 1980s. Unlike today's armies of hired tech lobbyists who line K Street and advocate on companies' behalf, early lobbying efforts involved the CEOs themselves, bringing a little California casual to Capitol Hill's marble hallways.

Semiconductor executives descended on Washington to plead for protection from the heavily subsidized Japanese competitors who were flooding, and dominating, the chip market. Venture capitalists and electronics executives successfully petitioned for reductions in the capital gains taxes that, in their telling, scared investors away from making "risk capital" bets on new tech companies.

Even the proud nonvoter Steve Jobs spent two weeks on the Hill in 1982 attempting to persuade lawmakers to pass a tax break that would encourage widespread use of computers in schools—an education market that, conveniently, Apple already dominated. The legislation failed in Washington, but then-Governor Jerry Brown of California swooped in to sign a similar state-level measure, thereby helping ensure an Apple would be the first computer for most of the schoolchildren in the nation's most populous state.

The well-trod path between Silicon Valley and Washington became even wider in the 1990s, as the commercialization of the Internet spurred a new set of technologists to collaborate with lawmakers in an effort to influence the rules and regulations that would guide the new platform. Silicon Valley dot-com companies and their allies then were the David to the Goliath of the telecom industry, and a main goal of their engagement in the legislative debates around telecommunication reform and Internet regulation was to ensure that online information flowed as freely as possible—without the Comcasts and AT&Ts of the world choking it off.

In a divided Washington, the information-should-be-free message was one with appeal to both the Democrats in the Clinton-Gore White House and the Republicans in the Gingrich-led Congress. The result was that the Internet industry was allowed to regulate itself—precipitating an explosion in online content, and enabling the rise of new platforms and products that now form the core of tech's most significant companies.

So when it comes to tech and politics, there *is* a story there—a history urgently relevant to the technology landscape of 2020. The long history of defense work, and of companies that were both federal contractors and merchant producers, helps make legible the remarkable role that tech giants now play in national security and defense contracting. The consistent lobbying for tax and regulatory relief belies the Silicon Valley mythos that its innovative story was something that happened on its own, without government help, and it extends the political history of tech beyond only the military-industrial complex. The regulatory environment that governs the Internet today was the right response to the Silicon Valley of the early dot-com era, but it now is as outdated as pop-up ads and Windows 95.

Linking the two histories together, and appreciating their relevance, is essential grounding for where tech, and we who make and use it, might go next.

Margaret O'Mara (momara@uw.edu) is Howard and Frances Keller Endowed Professor in the Department of History, University of Washington, Seattle, WA, USA.

[a] Fred Turner, *From Counterculture to Cyberculture: Stewart Brand, The Whole Earth Catalog, and the Rise of Digital Utopianism.* University of Chicago Press, Chicago, 2006; Langdon Winner, "Cyberlibertarian Myths and the Prospects for Community," *Computers and Society 27*, 3 (Sept. 1997), 14–19; Richard Barbrook and Andy Cameron, "The Californian Ideology," *Mute 1*, 3 (Sept. 1, 1995); https://bit.ly/37p8Wt5

DOI:10.1145/3387107

Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang

Viewpoint
Federated Learning for Privacy-Preserving AI

Engineering and algorithmic framework to ensure data privacy and user confidentiality.

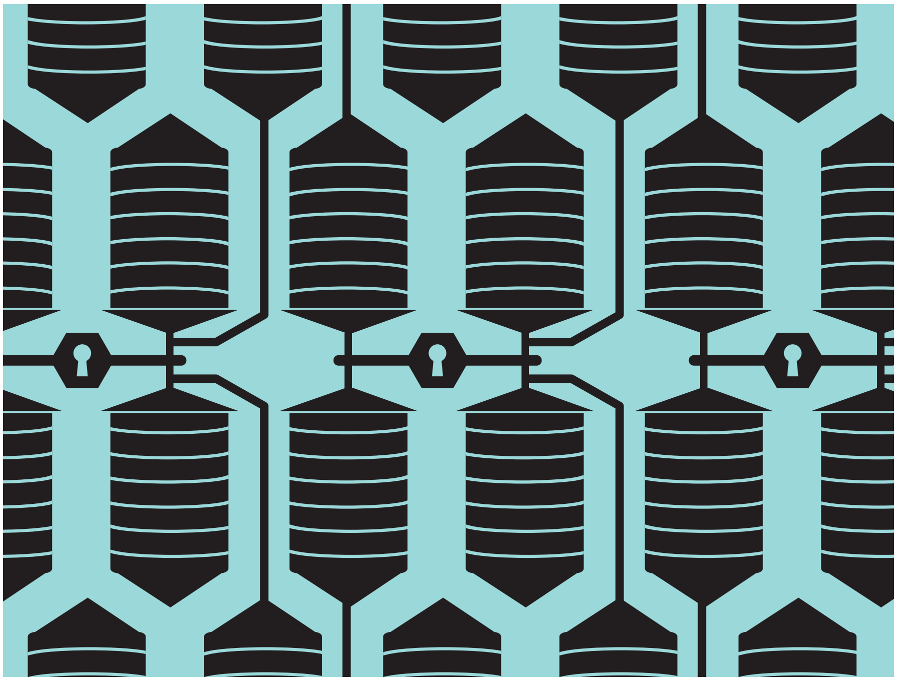


IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

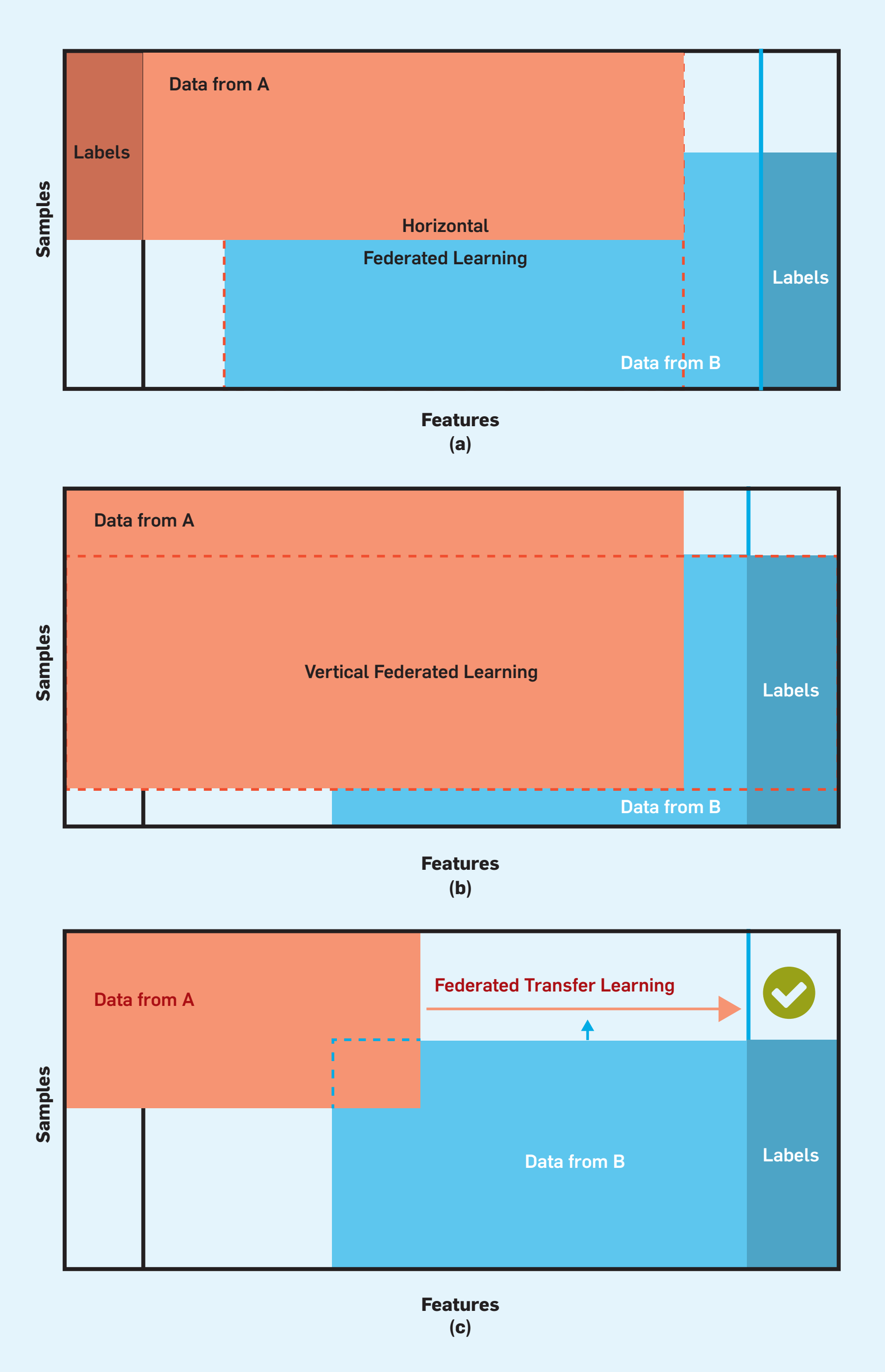
THERE HAS BEEN remarkable success of machine learning (ML) technologies in empowering practical artificial intelligence (AI) applications, such as automatic speech recognition and computer vision. However, we are facing two major challenges in adopting AI today. One is that data in most industries exist in the form of isolated islands. The other is the ever-increasing demand for privacy-preserving AI. Conventional AI approaches based on centralized data collection cannot meet these challenges. How to solve the problem of data fragmentation and isolation while complying with the privacy-protection laws and regulations is a major challenge for AI researchers and practitioners.

On the legal front, lawmakers and regulatory bodies are coming up with new laws ruling how data shall be managed and used.[3] One prominent example is the adoption of the General Data Protection Regulation (GDPR) by the European Union in 2018. In the United States, the California Consumer Privacy Act will be enacted in 2020. China's Cyber Security Law, came into effect in 2017, also imposed strict controls on data collection and transactions.

Under this new legislative landscape, collecting and sharing data among different organizations are becoming increasingly difficult, if not outright impossible. In addition, the sensitivity nature of certain data (for example, financial transactions and medical records) prohibits free data circulation and forces the data to exist in data silos. Due to competition, user privacy, data security, and complicated administrative procedures, even data integration among different departments of the same company faces heavy resistance. As the old privacy-intrusive way of collecting and sharing data are no longer allowed, data consolidation involving different data owners is becoming extremely challenging.[10]

Data silos and privacy concerns are two of the most challenging impediments to the AI progresses. It is thus natural to seek solutions to build ML models that do not rely on collecting data to a centralized storage where model training takes place. One attractive idea is to train a sub-model at each location with only local data, and then let the parties at different sites communicate their respective sub-models in order to reach a consensus for a global model. In order to ensure user privacy and data confidentiality, the communication process is carefully engineered so that no site can reverse-engineer the private data of any other sites. In the meanwhile, the model is built as if the data sources were combined. This is the idea be-

Figure 1. Categorization of FL.[10]



hind "federated machine learning," or "federated learning (FL)" for short.[7,9]

Definition

FL was practiced by Google for next-word prediction on mobile devices.[2,7] Google's FL system serves as an example of a secure distributed learning environment for B2C (business to consumer) applications where all parties share the same data features and collaboratively train an ML model. Besides the B2C paradigm, the FL framework has been extended to support "cross-silos" scenarios and B2B (business-to-business) applications by the AI researchers in WeBank,[a] where each party has different sets of data features.[5,6,9]

In a nutshell, a fundamental change in algorithmic design with FL is, instead of transferring raw data from sites to sites or to a server, we transfer ML model parameters in a secure way,[2,10] so that parties cannot access the content of others' data. FL is an algorithmic framework for building ML models that can be characterized by the following features.

- There are two or more parties interested in jointly building a model.
- Each party holds some data that it can use for model training.
- In the model-training process, the data held by each party does not leave that party.
- The model can be transferred in part from one party to another under an encryption scheme, such that any party cannot reverse-engineer the data of other parties.
- The performance of the federated model is a good approximation of an ideal model built with centralized data.

Techniques for privacy-preserving ML have been extensively studied,[1] such as employing differential privacy (DP)[4] and secure multi-party computation.[10] DP involves in adding noise to the training data, or using generalization methods to obscure certain sensitive features until the third party cannot distinguish the individual, thereby making the data impossible to be restored to protect user privacy. However, DP still requires that the data is transmitted elsewhere and involves a trade-off between accuracy and privacy.

Categorization

FL can be classified into horizontal FL (HFL),[7] vertical FL (VFL),[9] and federated transfer learning (FTL),[6] according to how data is distributed among the participating parties in the feature and sample spaces.[10] Figures 1a–1c illustrate the three FL categories respectively for a two-party scenario.

HFL refers to the scenarios where the parties share overlapping data features, but differ in data samples. Different from HFL, VFL applies to the scenarios where the parties share overlapping data samples, but differ in data features. FTL is applicable for the scenarios when there is little overlapping in data samples and in features. We also refer to HFL as sample-partitioned FL,[10] or example-partitioned FL,[5] as in a matrix form, samples correspond to the rows and features correspond to the columns (see Figure 1a). HFL is carried out across different horizontal rows, that is, data is partitioned by samples. We also refer to VFL as feature-partitioned

[a] WeBank—China's first Internet-only bank, see https://bit.ly/3o508ym

FL,[5,10] as VFL is carried out across different vertical columns, that is data is partitioned by features (see Figure 1b).

For example, when two organizations providing different services (for example, a bank and an e-commerce company), but having a large intersection of common customers (that is, aligned data samples), they may collaborate on the different data features they respectively own to achieve better ML models using VFL.[5,9]

Architecture

An FL system architecture can employ the client-server model, as shown in Figure 2a. The coordinator C can be played by an authority such as a government department or replaced by a secure computing node.[10] The communication between the coordinator C and the data owners A and B (a.k.a. parties) may be encrypted (for example, using homomorphic encryption[2,10]) to further defend against any information leakage. Further, the coordinator C may also be a logical entity and be located in either A or B. An FL system architecture can also employ the peer-to-peer model, without a coordinator, as illustrated in Figure 2b. The data owners A and B communicate directly without the help of a third party. While there are only two data owners in Figure 2, an FL system may generally include two or more data owners.[7,10]

Taking the client-server model shown in Figure 2a as an example, we summarize the encrypted and secure model training with VFL into the following four steps, after aligning the data samples between the two data owners.[10]

- **Step 1:** C creates encryption key pairs, and sends the public key to A and B.
- **Step 2:** A and B encrypt and exchange intermediate computation results for gradient and loss calculations.
- **Step 3:** A and B compute encrypted gradients and add an additional mask, respectively. B also computes the encrypted loss. A and B send encrypted results to C.
- **Step 4:** C decrypts gradients and loss and sends the corresponding results back to A and B. A and B unmask the gradients, and update their model parameters accordingly.

Readers can find more more information about the FL model training and inference procedures, such as the convergence speeds of the exiting training algorithms, in the existing works[5,10] and references therein.

Application Examples

FL enables us to build cross-enterprise, cross-data and cross-domain AI applications while complying with data protection laws and regulations. It has potential applications in finance, insurance, healthcare, education, smart city, and edge computing, and so forth.[10] We present here two FL applications selected from the use cases[b] that have been deployed in practice by WeBank.

Use Case 1: FedRiskCtrl

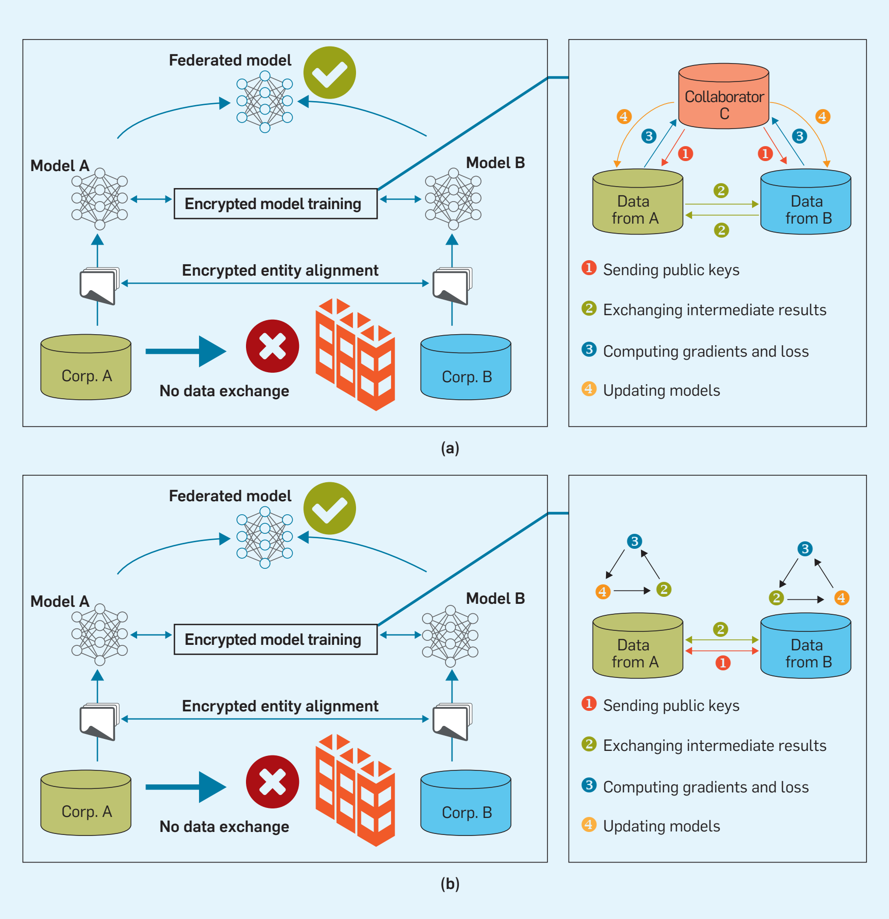
The first use case is an FL application in finance. It is an example of federated risk control (FedRiskCtrl) for small and micro enterprise (SME) loans deployed by WeBank.[c]

There is an invoice agency A, which has invoice related data features, such as $\left\{X_m^{(k)}\right\}_{m=1}^{M}$ for the k^{th} SME. There is bank B, which has credit-related data features, such as $\left\{X_n^{(k)}\right\}_{n=M+1}^{N}$ and label $Y^{(k)}$ for the k^{th} SME, with $N > M$. The agency A and bank B collaboratively build a risk control model for SME loans using VFL.[10]

Before model training, we need to find the common SMEs served by A and B to align the training data samples, which is called private set intersection or secure entity alignment.[10] After determining the aligned data samples between A and B, we can then follow the steps shown in Figure 2 for training a risk control model for SME loans.

FedRiskCtrl is implemented with the FATE (Federated AI Technology Enabler) platform.[d] With VFL, the agency A and the bank B do not need to expose their private data to each other, and the model built with FL is expected to perform as well as the

Figure 2. Examples of VFL Architecture.[10]



[b] For more use cases deployed by WeBank, see https://bit.ly/37lgjlq

[c] For more details on FedRiskCtrl, see https://bit.ly/3o8ECbY

[d] For more information on FATE, see https://bit.ly/37hBC7r

FL can overcome the challenges of data silos, small data, privacy issues, and lead us toward privacy-preserving AI.

model built with centralized dataset $\left\{\left\{X_i^{(k)}\right\}_{i=1}^{N}, Y^{(k)}, k=1,2,\ldots,K\right\}$. The model built with FL performs significantly better than the model built only with the bank B's data.

Use Case 2: FedVision

The second use case is an FL application in edge computing. It is an example of federated computer vision (FedVision) for object detection deployed by WeBank.[e]

Due to privacy concerns and high cost of transmitting video data, it is difficult to centrally collect surveillance video data for model training in practice. With FedVision, surveillance video data collected and stored in the edge cloud of each surveillance company are no longer required to be uploaded to a central cloud for centralized model training.[10] In FedVision, an initial object detection model is sent from the FL server to each surveillance company (that is, to each edge cloud), which then uses the locally stored data to train the object detection model. After a few local training epochs, the model parameters from each surveillance company are encrypted and sent to the FL server. The local model parameters are aggregated into a global federated model by the FL server and sent back to each surveillance company. This process iterates until the stopping criterion is met.

The model training process in FedVision is very similar to the federated averaging procedure for HFL model training.[2,7] The final global federated model will be distributed to the participating surveillance companies in the federation, to be used for object detection, such as fire detection.

Outlook

FL can overcome the challenges of data silos, small data, privacy issues, and lead us toward privacy-preserving AI. It will become the foundation of next-generation ML that caters to technological and societal needs for responsible AI development and applications.[10]

While FL has great potential, it also faces several practical challenges.[5,10] The communication links between the local data owners and the coordinator may be slow and unstable. There may be a very large number of local data owners (for example, mobile devices) to manage. Data from different data owner in an FL system may follow non-identical distributions, and different data owners may have unbalanced numbers of data samples, which may result in a biased model or even failure of model training.[5] Incentivizing mobile device owners or organizations to participate in FL also needs further studies. Incentive mechanism design for FL should be done in such a way to make the federation fair and sustainable.[8] C

[e] For more details on FedVision, see https://bit.ly/3o6runW

References

1. Al-Rubaie, M., and Chang, J.M. Privacy-preserving machine learning: Threats and solutions. *IEEE Security and Privacy* (Apr. 2019).
2. Bonawitz, K. et al. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of ACM SIGSAC CCS'17* (Nov. 2017).
3. DLA Piper. Data protection laws of the world: Full handbook (Jan. 2020); https://bit.ly/354nDiC
4. Dwork, C. Differential privacy: A survey of results. In *Proceedings of TAMC'08* (Apr. 2008).
5. Kairouz, P. et al. Advances and open problems in federated learning. (Dec. 2019); arXiv preprint arXiv:1912.04977
6. Liu, Y., Chen, T., and Yang, Q. Secure federated transfer learning. In *Proceedings of IJCAI'19* (Aug. 2019).
7. McMahan, H.B., Moore, E., Ramage, D., and y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of AISTATS'17* (Apr. 2017).
8. Richardson, A., Filos-Ratsikas, A., and Faltings, B. Rewarding high-quality data via influence functions. (Aug. 2019); arXiv preprint arXiv:1908.11598
9. Yang, Q. et al. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol. (TIST)* (Feb. 2019).
10. Yang, Q. et al. *Federated Learning.* Morgan & Claypool, Dec. 2019.

Yong Cheng (petercheng@webank.com) is a Senior Researcher at WeBank, China.

Yang Liu (yangliu@webank.com) is a Senior Researcher at WeBank, China.

Tianjian Chen (tobycheng@webank.com) is a Deputy General Manager of AI at WeBank, China.

Qiang Yang (qyang@cse.ust.hk) is Chief AI Officer at WeBank, China, and a Chair Professor at The Hong Kong University of Science and Technology, China.

DOI:10.1145/3410625

Article development led by acmqueue
queue.acm.org

Be kind and rewind.

BY JESSIE FRAZELLE

The Life of a Data Byte

more online

A version of this article with embedded informational links is available at https://queue.acm.org/detail.cfm?id=3419941

A BYTE OF data has been stored in a number of different ways through the years as newer, better, and faster storage media are introduced. A byte is a unit of digital information that most commonly refers to eight bits. A bit is a unit of information that can be expressed as 0 or 1, representing a logical state. Let's take a brief walk down memory lane to learn about the origins of bits and bytes.

Going back in time to Babbage's Analytical Engine, you can see that a bit was stored as the position of a mechanical gear or lever. In the case of paper cards, a bit was stored as the presence or absence of a hole in the card at a specific place. For magnetic storage devices, such as tapes and disks, a bit is represented by the polarity of a certain area of the magnetic film. In modern DRAM (dynamic random-access memory), a bit is often represented as two levels of electrical charge stored in a capacitor, a device that stores electrical energy in an electric field. (In the early 1960s, the paper cards used to input programs for IBM mainframes were known as *Hollerith cards*, named after their inventor Herman Hollerith from the Tabulating Machines Company—which through numerous mergers is what is now known as IBM.)

In June 1956, Werner Buchholz coined the word *byte* to refer to a group of bits used to encode a single character of text. Let's address character encoding, starting with ASCII (American Standard Code for Information Interchange). ASCII was based on the English alphabet; therefore, every letter, digit, and symbol (a-z, A-Z, 0–9, +, -, /, ", !, among others) were represented as a seven-bit integer between 32 and 127. This wasn't very friendly to other languages. To support other languages, Unicode extended ASCII so that each character is represented as a code-point, or character; for example, a lowercase j is U+006A, where U stands for Unicode followed by a hexadecimal number.

UTF-8 is the standard for representing characters as eight bits, allowing every code-point from 0 to 127 to be stored in a single byte. This is fine for English characters, but other languages often have characters that are expressed as two or more bytes. UTF-16 is the standard for representing characters as 16 bits, and UTF-32 is the standard for 32 bits. In ASCII every character is a byte, but in Unicode, that's often not true—a character can be one, two, three, or more bytes. Groups of characters might also be referred to as *words*, as in this linked Univac ad calling out "1 kiloword or 12,000 characters." This article refers throughout to different sized groupings of bits—the number of bits in a byte varying according to the design of the storage medium in the past.

IMAGE BY ALESANKO RODRIGUEZ

This article also travels in time through various storage media, diving into how data has been stored throughout history. By no means does this include every single storage medium ever manufactured, sold, or distributed. This article is meant to be fun and informative but not encyclopedic. It wraps up with a look at the current and future technologies for storage.

To get started, let's assume we have a byte of data to be stored: the letter j, or as an encoded byte 6a, or in binary 01001010. As we travel through time, this data byte will come into play in some of the storage technologies covered here.

1951

The story begins in 1951 with the Uniservo tape drive for the Univac 1 computer, the first tape drive made for a commercial computer. The tape was three pounds of a thin strip (half-inch) of nickel-plated phosphor bronze, called vicalloy, which was 1,200 feet long. Our data byte could be stored at a rate of 7,200 characters per second on tape moving at 100 inches per second. At this point in history, you could measure the speed of a storage algorithm by the distance the tape traveled.

1952

Let's fast forward a year to May 21, 1952, when IBM announced its first magnetic tape unit, the IBM 726. Our data byte could now be moved off Uniservo metal tape onto IBM's magnetic tape. This new home would be super cozy for our very small data byte since the tape could store up to two million digits. This magnetic seven-track tape moved at 75 inches per second with a transfer rate of 12,500 digits or 7,500 characters—called *copy groups* at the time—per second. For reference, this article has 35,123 characters.

Seven-track tapes had six tracks for data and one to maintain parity by ensuring that the total number of 1-bits in the string was even or odd. Data was recorded at 100 bits per linear inch. This system used a *vacuum channel* method of keeping a loop of tape circulating between two points. This allowed the tape drive to start and stop the tape in a split second. This was done by placing long vacuum columns between the tape reels and the read/write heads to absorb sudden increases in tension in the tape, without which the tape would have typically broken. A removable plastic ring on the back of the tape reel provided write protection. About 1.1MB could be stored on one reel of tape.

Think back to VHS tapes: What was required before returning a movie to Blockbuster? Rewinding the tape! The same could be said for the tape used for computers. Programs could not hop around a tape, or randomly access data—they had to read and write in sequential order.

> **What was required before returning a movie to Blockbuster? Rewinding the tape! The same could be said for the tape used for computers. Programs could not hop around a tape, or randomly access data—they had to read and write in sequential order.**

1956

The era of magnetic-disk storage began in 1956 with IBM's completion of the 305 RAMAC computer delivered to Zellerbach Paper in San Francisco. This computer was the first to use a moving-head HDD (hard-disk drive). The RAMAC disk drive consisted of 50 magnetically coated 24-inch-diameter metal platters capable of storing about five million characters of data, seven bits per character, and spinning at 1,200 revolutions per minute. The storage capacity was about 3.75MB.

RAMAC allowed real-time random access memory to large amounts of data, unlike magnetic tape or punch cards. IBM advertised the RAMAC as being able to store the equivalent of 64,000 punched cards. Previously, transactions were held until a group of data was accumulated and batch processed. The RAMAC introduced the concept of continuously processing transactions as they occurred so data could be retrieved immediately when it was fresh. Our data byte could now be accessed in the RAMAC at 100,000 bits per second. Prior to this, with tapes, people had to write and read sequential data and could not randomly jump to various parts of the tape. Real-time random access of data was truly revolutionary at this time.

1963

DECtape was introduced in 1963. Its namesake was the Digital Equipment Corporation, known as DEC for short. DECtape was inexpensive and reliable so it was used in many generations of DEC computers. This ¾-inch tape was laminated and sandwiched between two layers of Mylar on a four-inch reel.

DECtape could be carried by hand, as opposed to its large weighty predecessors, making it great for personal computers. In contrast to seven-track tape, DECtape had six data tracks, two mark tracks, and two clock tracks. Data was recorded at 350 bits per inch. Our data byte, which is eight bits but could be expanded to 12, could be transferred to DECtape at 8,325 12-bit words per second with a tape speed of 93 +/-12 inches per second. This is 8% more digits per second than the Uniservo metal tape from 1952.

1967

Four years later in 1967 a small team at IBM started working on the IBM floppy-disk drive, code-named Minnow. At the time, the team was tasked with developing a reliable and inexpensive way to load microcode into IBM System/370 mainframes. The project then got reassigned and repurposed to load microcode into the controller for the IBM 3330 Direct Access Storage Facility, code-named Merlin.

Our data byte could now be stored on read-only eight-inch flexible Mylar disks coated with magnetic material, known as *floppy disks*. At the time of release, the result of the project was named the IBM 23FD Floppy Disk Drive System. The disks could hold 80KB of data. Unlike hard drives, a user could easily transfer a floppy in its protective jacket from one drive to another. Later, in 1973, IBM released a read/write floppy-disk drive, which then became an industry standard.

1969

In 1969, the AGC (Apollo Guidance Computer) read-only rope memory was launched into space aboard Apollo 11, which carried American astronauts to the moon and back. This rope memory was made by hand and could hold 72KB of data. Manufacturing rope memory was laborious, slow, and required skills analogous to textile work; it could take months to weave a program into the rope memory, illustrated in the accompanying figure. But it was the right tool for the job at the time to resist the harsh rigors of space. When a wire went through one of the circular cores, it represented a 1. Wires that went around a core represented a 0. Our data byte would take a human a few minutes to weave into the rope.

1977

The Commodore PET, the first (successful) mass-market personal computer, was released in 1977. Built in to the PET was a Commodore 1530 Datasette (a portmanteau of *data* plus *cassette*). The PET converted data into analog sound signals that were then stored on cassettes. This made for a cost-effective and reliable storage solution, albeit very slow. Our small data byte could be transferred at a rate of around 60–70 bytes per second. The cassettes could hold about 100KB per 30-minute side, with two sides per tape. For example, you could fit about two ("rickroll" warning) 55KB images on one side of the cassette. The Datasette also appeared in the Commodore VIC-20 and Commodore 64.

1978

Let's jump ahead a year to 1978 when the LaserDisc was introduced as Discovision by MCA and Philips. *Jaws* was the first film sold on a LaserDisc in North America. The audio and video quality on a LaserDisc was far better than the competitors', but too expensive for most consumers. As opposed to VHS tape, which consumers could use to record TV programs, the LaserDisc could not be written to. LaserDiscs used analog video with analog FM stereo sound and PCM (pulse-code modulation) digital audio. The disks were 12 inches in diameter and composed of two single-sided aluminum disks layered in plastic. The LaserDisc is remembered today as being the foundation CDs that DVDs were built upon.

1979

A year later in 1979 Alan Shugart and Finis Conner founded Seagate Technology with the idea of scaling down a hard-disk drive to be the same size as a 5¼-inch floppy disk, which at the time was the standard. Their first product, in 1980, was the Seagate ST506, the first HDD for microcomputers. The 5¼-inch disk held 5MB of data, which at the time was five times more than the standard floppy disk. It was a rigid, metallic platter coated on both sides with a thin layer of magnetic material to store data. Our data byte could be transferred at a speed of 625KB/s onto the disk. That's about one (second and final "rickroll" warning) 625KB animated GIF per second.

1981

A couple of years later Sony introduced the first 3½-inch floppy drive. Hewlett-Packard was the first adopter of the technology in 1982 with its HP-150. This put the 3½-inch floppy disk on the map and gave it wide distribution in the industry. The disks were single sided with a formatted capacity of 161.2KB and an unformatted capacity of 218.8KB. In 1982 the double-sided version was made available, and the Microfloppy Industry Committee, a consortium of 23 media companies, based a spec for a 3½-inch floppy on Sony's original designs, cementing the format into history. Our data byte could now be stored on the early version of one of the most widely distributed storage media: the 3½-inch floppy disk.

1984

In 1984, the CD-ROM (compact disc read-only memory), holding 550MB of prerecorded data, was announced by Sony and Philips. This format grew out of CD-DA (compact disc digital audio), developed by the two companies in 1982. The CD-DA, which was used for distributing music, had a capacity of 74 minutes. When Sony and Philips were negotiating the standard for a CD-DA, legend has it that one of the four people involved insisted it be able to hold *all of Beethoven's Ninth Symphony*. The first product released on CD-ROM was Grolier's Electronic Encyclopedia, which came out in 1985. The encyclopedia contained nine million words, occupying only 12% of the available space, which was 553 mebibytes. There would be more than enough room for the encyclopedia and our data byte. Shortly thereafter in 1985, computer and electronics

Rope memory.

FIGURE BY MARK RICHARDS/COMPUTER HISTORY MUSEUM

companies worked together to create a standard for the disks so any computer would be able to access the information.

Also in 1984, Fujio Masuoka published his work on a new type of floating-gate memory, called *flash memory*, which was capable of being erased and reprogrammed multiple times.

Let's first review how floating-gate memory works. It uses transistors, which are electrical gates that can be switched on and off individually. Since each transistor can be in two distinct states (on or off), it can store two different numbers: 0 and 1. *Floating gate* refers to the second gate added to the middle transistor. This second gate is insulated by a thin oxide layer. These transistors use a small voltage applied to the gate of the transistor to denote whether it is on or off, which in turn translates to a 0 or 1.

With a floating gate, when a suitable voltage is applied across the oxide layer, the electrons tunnel through it and get stuck on the floating gate. Therefore, even if the power is disconnected, the electrons remain present on the floating gate. When no electrons are on the floating gate, it represents a 1; and when electrons are trapped on the floating gate, it represents a 0. Reversing this process and applying a suitable voltage across the oxide layer in the opposite direction causes the electrons to tunnel off the floating gate and restore the transistor to its original state. Therefore, the cells are made programmable and nonvolatile. Our data byte could be programmed into the transistors as 01001010, with electrons trapped in the floating gates to represent the zeros.

Masuoka's design was a bit more affordable but less flexible than EEPROM (electrically erasable programmable read-only memory), since it required multiple groups of cells to be erased together, but this also accounted for its speed. At the time, Masuoka was working for Toshiba but quit the company shortly after to become a professor at Tohoku University. He was displeased with Toshiba for not rewarding him for his work and sued the company, demanding compensation for his work. The case settled in 2006 with a one-time payment of ¥87m, equivalent to $758,000. This still seems low, given how impactful flash memory has been on the industry.

While on the topic of flash memory, let's look at the difference between NOR and NAND flash. Flash stores information in memory cells made up of floating-gate transistors. The names of the technologies are tied directly to the way the memory cells are organized.

In NOR flash, individual memory cells are connected in parallel, allowing the random access. This architecture enables the short read times required for the random access of microprocessor instructions. NOR flash is ideal for lower-density applications that are mostly read only. This is why most CPUs typically load their firmware from NOR flash. Masuoka and colleagues presented the invention of NOR flash in 1984 and NAND flash in 1987.

In contrast, NAND flash designers gave up the ability for random access in a tradeoff to gain a smaller memory cell size. This also has the benefits of a smaller chip size and lower cost-per-bit. NAND flash's architecture consists of an array of eight memory transistors connected in a series. This leads to high-storage density, smaller memory-cell size, and faster write and erase since it can program blocks of data at a time. This comes at the cost of having to overwrite data when it is not sequentially written and data already exists in a block.

1991

Let's jump ahead to 1991 when a prototype SSD (solid-state disk) module was made for evaluation by IBM from SanDisk, at the time known as SunDisk. This design combined a flash storage array and nonvolatile memory chips with an intelligent controller to detect and correct defective cells automatically. The disk was 20MB in a 2½-inch form factor and sold for approximately $1,000. IBM wound up using it in the ThinkPad pen computer.

1994

In 1994, Iomega released the Zip disk, a 100MB cartridge in a 3½-inch form factor, a bit thicker than a standard 3½-inch disk. Later versions of the Zip disk could store up to 2GB. These disks had the convenience of being as small as a floppy disk but with the ability to hold a larger amount of data, which made them compelling. Our data byte could be written onto a Zip disk at 1.4MB/s. At the time, a 1.44MB 3½-inch floppy would write at about 16kB/s. In a Zip drive, heads are noncontact read/write and fly above the surface, which is similar to a hard drive but unlike other floppies. Because of reliability problems and the affordability of CDs, Zip disks eventually became obsolete.

Also in 1994, SanDisk introduced CompactFlash, which was widely adopted into consumer devices such as digital and video cameras. Like CD-ROMs, CompactFlash speed is based on *x*-ratings (8x, 20x, 133x, and so on). The maximum transfer rate is calculated based on the original audio CD transfer rate of 150kB/s. This winds up looking like R = K × 150kB/s, where R is the transfer rate and K is the speed rating. For 133x CompactFlash, our data byte would be written at 133 × 150kB/s or around 19,950kB/s or 19.95MB/s. The CompactFlash Association was founded in 1995 to create an industry standard for flash-based memory cards.

1997

A few years later in 1997, the CD-RW (compact disc rewritable) was introduced. This optical disc was used for data storage, as well as backing up and transferring files to various devices. CD-RWs can be rewritten only about 1,000 times, which at the time was not a limiting factor since users rarely overwrote data on one disk.

CD-RWs are based on phase-change technology. During a phase change of a given medium, certain properties of the medium change. In the case of CD-RWs, phase shifts in a special compound, composed of silver, tellurium, and indium, cause *reflecting lands* and *non-reflecting bumps*, each representing a 0 or 1. When the compound is in a crystalline state, it is translucent, which indicates a 1. When the compound is melted into an amorphous state, it becomes opaque and nonreflective, which indicates a 0. We could write our data byte 01001010 as non-reflecting bumps and reflecting lands this way.

CD-RWs eventually lost much of their market share to DVDs.

1999

In 1999, IBM introduced the smallest hard drives in the world at the time: the

IBM microdrive in 170MB and 340MB capacities. These were small hard disks, one inch in size, designed to fit into CompactFlash Type II slots. The intent was to create a device to be used like CompactFlash but with more storage capacity. These were soon replaced by USB flash drives, however, and larger CompactFlash cards once they became available. Like other hard drives, microdrives were mechanical and contained small, spinning disk platters.

2000

USB flash drives were introduced in 2000. These consisted of flash memory encased in a small form factor with a USB interface. Depending on the version of the USB interface used, the speed varies: USB 1.1 is limited to 1.5Mbps, whereas USB 2.0 can handle 35Mbps, and USB 3.0 can handle 625Mbps. The first USB 3.1 type C drives were announced in March 2015 and have read/write speeds of 530Mbps. Unlike floppy and optical disks, USB devices are harder to scratch but still deliver the same use cases of data storage and transferring and backing up files. Because of this, drives for floppy and optical disks have since faded in popularity, replaced by USB ports.

2005

HDD manufacturers started shipping products using PMR (perpendicular magnetic recording) in 2005. Interestingly, this happened at the same time that Apple announced the iPod Nano, which used flash as opposed to the one-inch hard drives in the iPod Mini, causing a bit of an industry hoohaw.

A typical hard drive contains one or more rigid disks coated with a magnetically sensitive film consisting of tiny magnetic grains. Data is recorded when a magnetic write-head flies just above the spinning disk, much like a record player and a record, except a needle is in physical contact with the record. As the platters spin, the air in contact with them creates a slight breeze. Just as air on an airplane wing generates lift, the air generates lift on the head's airfoil. The write-head rapidly flips the magnetization of one magnetic region of grains so that its magnetic pole points up or down to denote a 1 or a 0.

The predecessor to PMR was LMR (longitudinal magnetic recording). PMR can deliver more than three times the storage density of LMR. The key difference between the two is that the grain structure and the magnetic orientation of the stored data of PMR media is columnar instead of longitudinal. PMR has better thermal stability and improved SNR (signal-to-noise ratio) as a result of better grain separation and uniformity. It also benefits from better writability because of stronger head fields and better magnetic alignment of the media. Like LMR, PMR's fundamental limitations are based on the thermal stability of magnetically written bits of data and the need to have sufficient SNR to read back written information.

Storage class memory is persistent but goes further by also providing performance better than or comparable to primary memory, as well as byte addressability.

2007

Hitachi Global Storage Technologies announced the first 1TB HDD in 2007. The Hitachi Deskstar 7K1000 used five 3.5- inch 200GB platters and rotated at 7,200 RPM. This is in stark contrast to the world's first HDD, the IBM RAMAC 350, which had a storage capacity of approximately 3.75MB. How far we have come in 51 years! But wait, there's more.

2009

In 2009, technical work was beginning on NVMe (nonvolatile memory express). NVM is a type of memory that has persistence, in contrast to volatile memory, which needs constant power to retain data. NVMe filled a need for a scalable host controller interface for PCIe (peripheral component interconnect express)-based SSDs. More than 90 companies were part of the working group that developed the design. This was all based on prior work to define the NVMHCIS (nonvolatile memory host controller interface specification). Opening up a modern server would likely result in finding some NVMe drives. The best NVMe drives today can do about a 3,500MB/s read and 3,300MB/s write. For the data byte we started with, the character j, that is extremely fast compared with a couple of minutes to handweave rope memory for the Apollo Guidance Computer.

Today and the Future

Now that we have traveled through time a bit, let's take a look at the state of the art for SCM (storage class memory). Like NVM, SCM is persistent but goes further by also providing performance better than or comparable to primary memory, as well as byte addressability. SCM aims to address some of the problems faced by caches today such as the low density of SRAM (static random access memory). DRAM provides better density, but this comes at a cost of slower access times. DRAM also suffers from requiring constant power to refresh memory.

Let's break this down a bit. Power is required since the electric charge on the capacitors leaks off little by little; this means that without intervention, the data on the chip would soon be lost. To prevent this leakage, DRAM requires an external memory-refresh circuit that periodically rewrites the data in the capacitors, restoring them to their original charge.

To solve the problems with density and power leakage, a few SCM technologies are in development: PCM (phase-change memory), STT-RAM (spin-transfer torque random access memory), and ReRAM (resistive random access memory). One nice aspect of all these technologies is their ability to function as MLCs (multilevel cells). This means they can store more than one bit of information, compared with SLCs (single-level cells), which can store only one bit per memory cell, or element. Typically, a memory cell consists of one MOSFET (metal-oxide-semiconductor field-effect transistor). MLCs reduce the number of MOSFETs required to store the same amount of data as SLCs, making them denser or smaller to deliver the same amount of storage as technologies using SLCs. Let's go over how each of these SCM technologies work.

Phase-change memory. PCM is similar to phase change for CD-RWs, described earlier. Its phase-change material is typically GST, or GeSbTe (germanium-antimony-tellurium), which can exist in two different states: amorphous and crystalline. The amorphous state has a higher resistance, denoting a 0, than the crystalline state, denoting a 1. By assigning data values to intermediate resistances, PCM can be used to store multiple states as an MLC.

Spin-transfer torque random access memory. STT-RAM consists of two ferromagnetic, permanent magnetic layers separated by a dielectric, an insulator that can transmit electric force without conduction. It stores bits of data based on differences in magnetic directions. One magnetic layer, called the reference layer, has a fixed magnetic direction, while the other magnetic layer, called the free layer, has a magnetic direction that is controlled by passing current. For a 1, the magnetization direction of the two layers are aligned. For a 0, the two layers have opposing magnetic directions.

Shingled magnetic recording results in a much more complex writing process, since writing to one track winds up overwriting an adjacent track.

Resistive random access memory. A ReRAM cell consists of two metal electrodes separated by a metal-oxide layer. This is similar to Masuoka's original flash-memory design, where electrons would tunnel through the oxide layer and get stuck in the floating gate or vice versa. With ReRAM, however, the state of the cell is determined by the concentration of oxygen vacancy in the metal-oxide layer.

SCM downsides and upsides. While these SCM technologies are promising, they still have downsides. PCM and STT-RAM have high write latencies. PCM's latencies are 10 times that of DRAM, while STT-RAM has 10 times the latencies of SRAM. PCM and ReRAM have a limit on write endurance before a hard error occurs, meaning a memory element gets stuck at a particular value.

In August 2015, Intel announced Optane, a product built on 3DXPoint, pronounced 3D cross-point. Optane claims performance 1,000 times faster than NAND SSDs with 1,000 times the performance, while being four to five times the price of flash memory. Optane is proof that SCM is not just experimental. It will be interesting to watch how these technologies evolve.

Helium hard-disk drive. An HHDD is a high-capacity HDD filled with helium and hermetically sealed during manufacturing. Like other hard disks, covered earlier, it is much like a record player with a rotating magnetic-coated platter. Typical HDDs would just have air inside the cavity, but that air causes an amount of drag on the spin of the platters.

Helium balloons float, so we know helium is lighter than air. Helium is, in fact, one-seventh the density of air, therefore reducing the amount of drag on the spin of the platters, causing a reduction in the amount of energy required for the disks to spin. This is actually a secondary feature; the primary feature of helium is to allow for packing seven platters in the same form factor that would typically hold only five. Attempting this with air-filled drives would cause turbulence. If you recall the airplane-wing analogy from earlier, this ties in perfectly. Helium reduces drag, thus eliminating turbulence.

As everyone knows, however, helium-filled balloons start to sink after a few days because the gas is escaping the balloons. The same could be said for these drives. It took years before manufacturers had created a container that prevented the helium from escaping the form factor for the life of the drive. Backblaze found that HHDDs had a lower annualized error rate of 1.03%, while standard hard drives resulted in 1.06%. Of course, that is so small a difference, it is hard to conclude much from it.

A helium-filled form factor can have an HDD encapsulated that uses PMR, or it could contain an MAMR (microwave-assisted magnetic recording) or HAMR (heat-assisted magnetic recording) drive. Any magnetic storage technology can be paired with helium instead of air. In 2014, HGST (Hitachi Global Storage Technologies) combined two cutting-edge technologies into its 10TB HHDD that used *host-managed SMR* (shingled magnetic recording).

Shingled magnetic recording. As noted earlier, PMR was SMR's predecessor. In contrast to PMR, SMR writes new tracks that overlap part of the previously written magnetic track, which in turn makes the previous track narrower, allowing for higher track density. The technology's name stems from the fact that the overlapping tracks are much like that of roof shingles.

SMR results in a much more complex writing process, since writing to one track winds up overwriting an adjacent track. This doesn't come into play when a disk platter is empty and data is sequential. Once you are writing to a series of tracks that already contain data, however, this process is destructive to existing adjacent data. If an adjacent track contains valid data, then it must be rewritten. This is quite similar to NAND flash, as covered earlier.

Device-managed SMR devices hide this complexity by having the device firmware manage it, resulting in an interface like any other hard disk you might encounter. *Host-managed* SMR devices, on the other hand, rely on the operating system to know how to handle the complexity of the drive.

Seagate started shipping SMR drives in 2013, claiming a 25% greater density than PMR.

Microwave-assisted magnetic recording. MAMR is an energy-assisted magnetic storage technology—like HAMR, which is covered next—that uses 20- to 40-GHz frequencies to bombard the disk platter with a circular microwave field. This lowers its coercivity, meaning that the magnetic material of the platter has a lower resistance to changes in magnetization. As already discussed, changes in magnetization of a region of the platter are used to denote a 0 or a 1, so since the platter has a lower resistance to changes in magnetization, the data can be written more densely. The core of this new technology is the spin torque oscillator used to generate the microwave field without sacrificing reliability.

Western Digital, also known as WD, unveiled this technology in 2017. Toshiba followed shortly after in 2018. While WD and Toshiba are busy pursuing MAMR, Seagate is betting on HAMR.

Heat-assisted magnetic recording. HAMR is an energy-assisted magnetic storage technology for greatly increasing the amount of data that can be stored on a magnetic device, such as an HDD, by using heat delivered by a laser to help write data onto the surface of a platter. The heat causes the data bits to be much closer together on the platter, which allows greater data density and capacity.

This technology is quite difficult to achieve. A 200-milliwatt laser heats a teensy area of the region to 750 degrees F (400 degrees C) quickly before writing the data, while not interfering with or corrupting the rest of the data on the disk. The process of heating, writing the data, and cooling must be completed in less than a nanosecond. These challenges required the development of nanoscale surface plasmons, also known as a surface-guided laser, instead of direct laser-based heating, as well as new types of glass platters and heat-control coatings to tolerate rapid spot-heating without damaging the recording head or any nearby data, and various other technical challenges that needed to be overcome.

Seagate first demonstrated this technology, despite many skeptics, in 2013. It started shipping the first drives in 2018.

more online

A version of this article with embedded informational links is available at https://queue.acm.org/detail.cfm?id=3419941

End Of Tape, Rewind

This article started off with the state of the art in storage media in 1951 and concludes after looking at the future of storage technology. Storage has changed a lot over time—from paper tape to metal tape, magnetic tape, rope memory, spinning disks, optical disks, flash, and others. Progress has led to faster, smaller, and more performant devices for storing data.

Comparing NVMe to the 1951 Uniservo metal tape shows that NVMe can read 486,111% more digits per second. Comparing NVMe to the Zip disks of 1994, you see that NVMe can read 213,623% more digits per second.

One thing that remains true is the storing of 0s and 1s. The means by which that is done vary greatly. I hope the next time you burn a CD-RW with a mix of songs for a friend, or store home videos in an optical disc archive (yup, you heard that right), you think about how the nonreflective bumps translate to a 0 and the reflective lands of the disk translate to a 1. If you are creating a mixtape on a cassette, remember that those are closely related to the Datasette used in the Commodore PET. Lastly, remember to be kind and rewind (this is a tribute to Blockbuster, but there are still open formats for using tape today).

Acknowledgments. Thank you to Robert Mustacchi, Trammel Hudson, and Rick Altherr for tidbits (I can't help myself) throughout this article.

Related articles on queue.acm.org

The Most Expensive One-byte Mistake
Poul-Henning Kamp
https://queue.acm.org/detail.cfm?id=2010365

Should You Upload or Ship Big Data to the Cloud?
Sachin Date
https://queue.acm.org/detail.cfm?id=2933408

Injecting Errors for Fun and Profit
Steve Chessin
https://queue.acm.org/detail.cfm?id=1839574

Jessie Frazelle is the cofounder and chief product officer of the Oxide Computer Company. Before that, she worked on various parts of Linux, including containers as well as the Go programming language.

DOI:10.1145/3424260

Article development led by acmqueue
queue.acm.org

The challenges of multifactor authentication based on SMS, including cellular security deficiencies, SS7 exploits, and SIM swapping.

BY ROGER PIQUERAS JOVER

Security Analysis of SMS as a Second Factor of Authentication

THE INTERNET STARTED becoming the main reason that household PSTN (public switched telephone network) lines stayed busy sometime in the mid to late 1990s. Over the next decade, a number of online services completely changed the interaction of society with technology. From email to the dawn of e-commerce, these services increasingly tied people to technology and the Internet.

Although the concept of a password was prevalent in many technology disciplines and domains, the general public had little knowledge of it, with the exception of PINs for their debit cards. It was the Internet that introduced the concept of password security to them. From that point on, people realized they had to remember passwords to access their email accounts, favorite e-commerce sites, and so on.

At that time, a password was all it took to unlock an account, and password requirements were very loose. The cybersecurity landscape was nowhere near as challenging as it is now, particularly when it comes to the security of consumer accounts. There were exceptions for certain industries, such as banking, where password requirements were slightly more stringent and a hidden form of two-factor authentication, based mainly on IP geolocation, was used before it became an option for other sites. Nevertheless, the average consumer generally needed just a simple password to access even highly critical repositories of data, and the same password was often reused for multiple accounts.

Today Internet security requires much more attention. A good example of what can go wrong is the hacking ordeal detailed by a technology reporter for *Wired* who was not using two-factor authentication for his email account.[8] Email accounts have become, over the years, not only large repositories of highly sensitive and private data, but also single points of failure for digital footprints on the Internet. For example, the majority of online services allow for the resetting of a password by sending an email to the user's main email account. As a result, if an email account gets compromised, many other accounts can also be compromised in short order.

As the security-threat landscape has changed, so too have the way passwords are used and their complexity requirements. Although many online services did not really adhere to best practices, it became widely acknowledged that passwords should be highly complex in order to maximize their entropy and, thus, substantially increase the amount of time it would take to crack them.

Eventually, though, some scientific studies[12,29] and a viral online cartoon[30] argued that increasing password complexity was not the right solution. Passphrases are proven to have much higher entropy and are much easier to remember. On the other hand, forcing password rotation, in combination with a strict password complexity policy, has

IMAGERY FROM SHUTTERSTOCK.COM

been shown to result in much weaker passwords. Moreover, as a rule of thumb, it is now acknowledged that a password might not necessarily have to be rotated if it is not present in any of the public repositories of leaked credentials in the wild.[10]

Along with the ongoing challenge of making passwords secure but still usable and easy to remember, the security industry recognized the security of an online account should not be protected only by something you know (your password). Somewhat similar to the banking approach, which requires something the user *has* (for example, a debit card) and something the user *knows* (for example, the card's PIN), online accounts started to support, and in some cases mandate, the use of two-factor authentication. The second factor needed to be something the user has—the obvious and simple choice was clear from the beginning: the user's smartphone.

Enabling two-factor authentication for online accounts is critical to their security. Everyone should enable this feature in (at the very least) their email accounts, as well as other accounts that store critical and sensitive data such as credit card numbers. Cryptocurrency exchange accounts, which are commonly the target of cybercriminals, should also be secured by multiple forms of authentication. The potentially high monetary value of what these accounts protect makes them an interesting case study of what might be the best choice for a second form of authentication. For example, while SMS (short message service) as a second form of authentication is a good idea for certain types of online accounts, it is not the best option for those who own a large amount of cryptocurrency in an online exchange.[3]

SMS-based authentication tokens are popular options for securing online accounts, and they are certainly more secure than using a password alone. The history of cellular network security, however, indicates that SMS is not a secure method of communication. From rogue base stations and stingrays to more sophisticated attacks, there are a number of known methods to eavesdrop on and brute-force text messages, both locally and remotely. As such, this method is not the most reliable for ac-

Figure 1. Pros and cons of different types of token.

Authentication Type	Advantages	Disadvantages
Application	► Convenient. ► No network connectivity required. ► The same application can be used to generate tokens for multiple accounts.	► Critical to generate and keep backup codes. ► Phone loss or theft. ► Cryptographic keys not always included in the backup of the phone.
SMS	► Not tied to cryptographic keys on the device. ► Easier to recover from device loss or theft.	► Requires network connectivity. ► Generally insecure.

Figure 2. Three methods of SMS interpretation.

Method of SMS Interception	Advantages	Disadvantages	Cost
Over the air	► Fast. ► Does not require any time-consuming preparation steps.	► Attacker must be in the vicinity of the victim. ► Technical knowledge of GSM, SW-radios and SW-based network stack.	► Low
SS7	► Can be done remotely. ► No need to interact with the operator or leave any trail. ► Fast. ► Does not require any time-consuming preparation steps	► Requires access to SS7 nodes and knowledge of SS7 protocols.	► High. ► For-sale access to SS7 nodes on the dark web.
SIM swap	► Can be done remotely. ► Low cost. ► Low-hanging fruit.	► Requires attacker to interact with network operator's customer support, which can be a time-consuming preparation stage. ► Leaves a trace. ► Some operators in African nations have mitigated this method.	► Low

counts that store assets with a high financial value, such as cryptocurrencies.

This article provides some insight into the security challenges of SMS-based multifactor authentication: mainly cellular security deficiencies, exploits in the SS7 (Signaling System No. 7) protocol, and the dangerously simple yet highly efficient fraud method known as SIM (subscriber identity module) swapping. Based on these insights, readers can gauge whether SMS tokens should be used for their online accounts. This article is not an actual analysis of multifactor authentication methods and what can be considered a second (or third, fourth, and so on) factor of authentication; for such a discussion, the author recommends reading security expert Troy Hunt's report on the topic.[9]

(Full disclosure: the author uses SMS to secure some rather vanilla online accounts, mainly those that do not require storing a credit card number or other sensitive financial information.)

SMS vs. One-Time Token App

For standard consumer online accounts, the two main options for providing a second factor of authentication are generally via SMS or leveraging a one-time token generated by an app on the user's smartphone. The latter is more secure and should be used for highly secure and sensitive accounts, but the former is the most widely used option and could be a valid choice in certain circumstances. Aside from their security, however, these two options have very different advantages and disadvantages in the context of convenience and usability—important factors to consider when designing a secure system. The pros and cons of both types of authentication are summarized in Figure 1.

App-generated token. As noted throughout this article, a one-time token generated by an application on the user's device is the most secure way of implementing two-factor authentication for online accounts without requiring nonstandard hardware to be used by the consumer (for example, an RSA token, a YubiKey, and so on, which are more common in an enterprise context). Aside from that, there are a number of advantages and disadvantages.

One of the main considerations for either option is network connectivity. The convenience of an app-generated token, which requires no network connectivity, contrasts with the strict requirement of connectivity to receive a token via SMS. Although network connectivity is considered a ubiquitous commodity, there are a number of scenarios in which a user could require access to an account while out of range.

Another advantage of app-based token delivery is that these apps can generally be registered and used with multiple online accounts. The main usability challenge with generating tokens with a smartphone app, however, is that administering such an app—and the cryptographic material it leverages—requires some extra effort. In general, backing up a smartphone to the cloud, the most common method, does not save such cryptographic material as part of the backup data. Nor does this material get saved on an unencrypted local backup on a computer. Even when locking a local backup with a password, not all seeds are stored with it. This can lead to users getting locked out of their accounts if their smartphones are lost or stolen, for example, or even if they get a new phone. In these scenarios, the so-called "backup codes" are important. As a rule of thumb, users should never wipe their old smartphones until the new ones have been fully set up and two-factor authentication apps have been reset.

SMS token. Two-factor authentication tokens received via text message tend to work well for standard consumer use because they're easy for the user.

There is no requirement to install an application on the user's device, and it doesn't require any management of backup codes or a backup plan to deal with a lost or stolen device. When a user gets a new device, there is no need to reset the two-factor authentication system, as text messaging is tied to the phone number, which generally remains unchanged on a new device.

On the downside, SMS-based authentication requires an active connection to the cellular network. Even though the majority of text message-based communication occurs over IP (for example, iMessage and WhatsApp), SMS second-factor authentication tokens are generally delivered over cellular networks' standard SMS. Therefore, Wi-Fi connectivity alone is not sufficient; an active cellular connection is necessary. This can be challenging in certain situations where cell service is spotty or nonexistent or connectivity is constrained to 802.11 networks.

Despite their security challenges, SMS-based authentication tokens are a widely utilized option, which is currently receiving active support from device manufacturers. As an example, Apple recently announced a new feature in iOS 14 to harden SMS codes against applications attempting to trick the user into inputting the code in a malicious app (https://developer.apple.com/news/?id=z0i801mg).

Security Challenges of SMS-Based Authentication

Despite its convenience and use by a large number of online services, two-factor authentication via text message has significant security challenges. This section presents an overview of the main security challenges of using SMS for two-factor authentication token delivery. These range from somewhat sophisticated threats to cellular network protocols, which require an adversary to be in the vicinity of the target victim, to low-hanging fruit techniques that, despite being much less technically complex, have no range constraints and can be implemented at near-zero cost. For example, one of the biggest security threats in mobile communication systems is SIM swapping, a systemic problem related to how mobile operators authenticate users in their customer care platforms.[18]

The main advantages and challenges of three different methods of SMS interception are summarized in Figure 2.

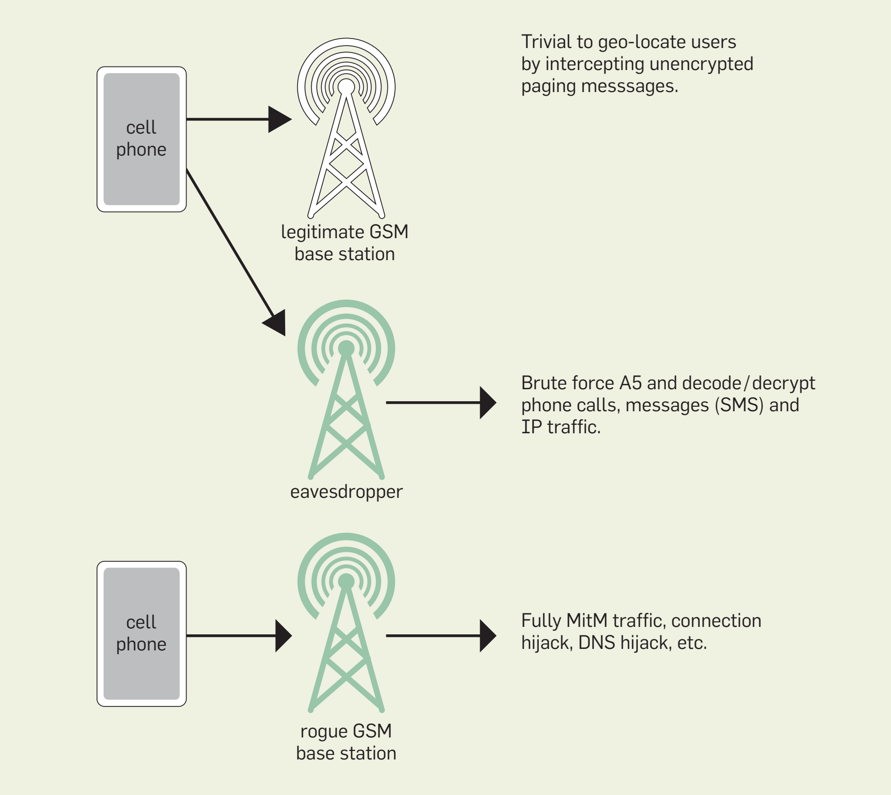
Cellular network security. The first generation of mobile networks (1G) lacked support for encryption. Legacy 2G GSM (Global System for Mobile Communications) networks lack mutual authentication and implement an outdated encryption algorithm. Combined with the wide availability of open source implementations of the GSM protocol stack, this resulted in the discovery of many possible exploits on the GSM radio link over the past decade[23] (illustrated in Figure 3). Specifically, both the techniques and the tools necessary to deploy a malicious GSM base station and implement a full MITM (man-in-the-middle) attack against a GSM connection are commodities, although they require the adversary to be in physical proximity to a given target. Low-cost software radios and open source implementations of the GSM protocol stack can be used to intercept mobile traffic, including SMS messages.[22,24]

Over the past few years, researchers have also demonstrated how to intercept SMS traffic with less strict proximity constraints by triggering a race condition when replying to paging messages intended for another user.[6] In order to geolocate a given target victim to intercept a token over SMS, privacy and location leaks have also been investigated in the context of legacy GSM networks.[17]

Specific efforts were made to enhance confidentiality and authentication in 3G and LTE (Long-term Evolution) mobile networks, with stronger cryptographic algorithms and mutual authentication implemented in both standards. Because of this, LTE has generally been considered secure, given its mutual authentication and strong encryption scheme. As such, confidentiality and authentication were wrongly assumed to be sufficiently guaranteed. Researchers demonstrated a few years ago, however, that LTE mobile networks are still vulnerable to protocol exploits, location leaks, and rogue base stations.[14,27]

Despite the strong cryptographic protection of user traffic and mutual authentication, a large number of control-plane (signaling) messages are regularly exchanged over an LTE radio link in the clear. Before the authentication and encryption steps of a connection are executed, a mobile device engages in a substantial conversation with any LTE base station (real or rogue)

Figure 3. GSM traffic eavesdropping and interception.


cell phone
legitimate GSM base station
Trivial to geo-locate users by intercepting unencrypted paging messsages.
eavesdropper
Brute force A5 and decode/decrypt phone calls, messages (SMS) and IP traffic.
cell phone
rogue GSM base station
Fully MitM traffic, connection hijack, DNS hijack, etc.

that advertises itself with the correct broadcast information. This results in a critical threat caused by the implicit trust placed—from the mobile device's point of view—in the messages coming from the base station. Many operations with critical security implications are executed when triggered by some of these implicitly trusted messages, which are neither authenticated nor validated.

In the age of large-scale cyberattacks, one of the largest civilian communication systems must rely on privacy protocols far more sophisticated than just basic implicit trust anchored in the base station looking like a legitimate station. Note that the same applies in reverse, with the base station implicitly trusting all preauthentication messages coming from mobile devices.

Although a malicious LTE base station is incapable of launching a full MITM attack, several studies have demonstrated and prototyped techniques to silently downgrade a modern smartphone to a vulnerable GSM connection.[13,26,27] What all these techniques have in common is they leverage and abuse such preauthentication messages.

As the industry prepares to embrace the advent of 5G, the security architecture for such next-generation mobile networks is being put under scrutiny. Several studies over the past year have highlighted the fact that most preauthentication message-based protocol exploits in LTE still apply to 5G networks.[11,15,16,26] As a result, silently downgrading the connection of a smartphone to a GSM link is still possible, given the current specifications for such mobile communication systems.[28] By abusing these vulnerabilities, an adversary could successfully intercept a two-factor authentication token delivered over SMS.

It is important to note, however, that intercepting tokens from SMS messages by intercepting GSM traffic is the most technologically complex option. Even though such attacks can be carried out with low-cost software radios and minor modifications to open source tools, the vast majority of fraud conducted by intercepting authentication tokens delivered via SMS leverages vulnerabilities in either SS7 or SIM swapping.

SS7 security. SS7 is a legacy architecture and protocol developed more than 30 years ago. It performs out-of-band signaling support for a number of functions in the PSTN, namely call establishment, billing, routing, and information exchange.[25] From its inception in 1988, when mobile operators started leveraging it for out-of-band signaling, this protocol's security mostly relied on the implicit trust among operators. It was regarded as a closed trusted network and had limited to no authentication built in. As a result, the security features of this network and protocol were minimal and depended on a small number of operators globally that were either state-controlled or large corporations. This is not the case anymore, as the number of operators is much larger as a result of the steep increase in mobile usage, as well as the growth in the number of MVNOs (mobile virtual network operators) around the globe over the past decade.

In the age of large-scale cyberattacks, one of the largest civilian communication systems must rely on privacy protocols far more sophisticated than just basic implicit trust anchored in the base station looking like a legitimate station.

3GPP (3rd Generation Partnership Project) added two new protocols to SS7 in the 1990s and 2000s: MAP (Mobile Application Part) and CAMEL (Customized Applications for Mobile Networks Enhanced Logic). These were aimed at supporting some of the new services that mobile networks provide, as well as new features for mobile operators.[19] Among other functionality, CAMEL allows the implementation of carrier-grade value-added services such as fraud control and prepaid services. In parallel, MAP provides services to geolocate devices globally, such as the anyTimeInterrogation service and LCS (Location Service). Discouragingly, none of these new SS7 subprotocols added authentication or security features.

Researchers have identified a number of critical security vulnerabilities in SS7 that could be exploited to geolocate users and intercept their traffic from nearly anywhere.[5] In some cases, the only requirement is to have access to the SS7 network, which, despite being more restricted now than in the past, can still easily be purchased on the dark web. Researchers have also gained access to the SS7 network via hacked femtocells and, in some scarce cases, actually purchasing access from mobile operators.

To make matters worse, researchers were also able to demonstrate techniques to intercept phone calls and text messages remotely by means of exploiting flaws in the CAMEL protocol. Figure 4 illustrates this process. Such security threats in mobile communication networks came to wide public attention when a German researcher demonstrated these attacks in a news report on primetime TV.[1]

Once the attacker has access to an entry point to the SS7 network, all it takes is one message to modify the registration for a given target in the MSC (mobile switching center), in the case of GSM. From that moment on, the MSC will reach out to the attacker instead.

Modern LTE networks largely migrated most SS7-based services over to the Diameter protocol. This new protocol, despite providing some improvements, still suffers from a number of vulnerabilities, most of which are flaws inherited from SS7.[2] As a result, similar remote interception of calls and text messages could be possible in LTE. Regardless, as discussed earlier, silently downgrading a smartphone to a much less secure GSM connection is simple, given enough proximity to the target.

Exploiting security flaws in SS7 networks and their protocols is a fairly efficient way to intercept two-factor authentication tokens delivered over SMS. In general, this is an attack vector that is known for being used by hacker groups and has become so relevant that it is even considered within the MITRE ATT&CK framework, which has been widely adopted by many technology companies as part of their security postures.[20]

SIM swapping. Despite the effectiveness of the SMS interception techniques that exploit flaws in cellular network protocols and in legacy SS7 networks, SIM swapping is arguably the number-one security threat against SMS communications being used to deliver one-time tokens for multifactor authentication.

As illustrated in Figure 5, a SIM swap attack consists of fooling a mobile operator, usually over a phone call with customer service, that the legitimate owner of a cellular subscription needs the account to be ported to a new SIM card. The caller may claim, for example, that the phone was lost overseas and access needs to be recovered as soon as possible on a newly acquired phone and a new SIM card. The credibility of such a story is actually not that important; SIM porting attacks are frustratingly easy to accomplish.[18]

Once an attacker successfully manages to get a victim's account ported to a SIM, the rest of the attack is fairly simple. From that moment until the victim notices the loss of coverage and calls customer service, the attacker will be the destination of any call and text message routed to the victim's MSISDN (mobile station international subscriber directory number)—that is, the victim's 10-digit phone number. Therefore, any two-factor authentication token requested will be received by the attacker.

This type of attack is simple to implement and accounts for a majority of breaches that require intercepting authentication tokens. Given the low cost and low effort a SIM swap attack requires, fraud and scamming rings are devising more sophisticated methods to scale up the number of accounts they can take over. For example, it was recently discovered that SIM swappers were bribing customer service employees to perform the swaps for them and even leveraging malware that targets the remote desktop technology used in call centers.[4]

Interestingly, fraudulent online account takeovers based on a SIM swap attack are not too complex to mitigate. Despite being a widely acknowledged

Figure 4. Exploiting flaws in the Camel SS7 protocol.

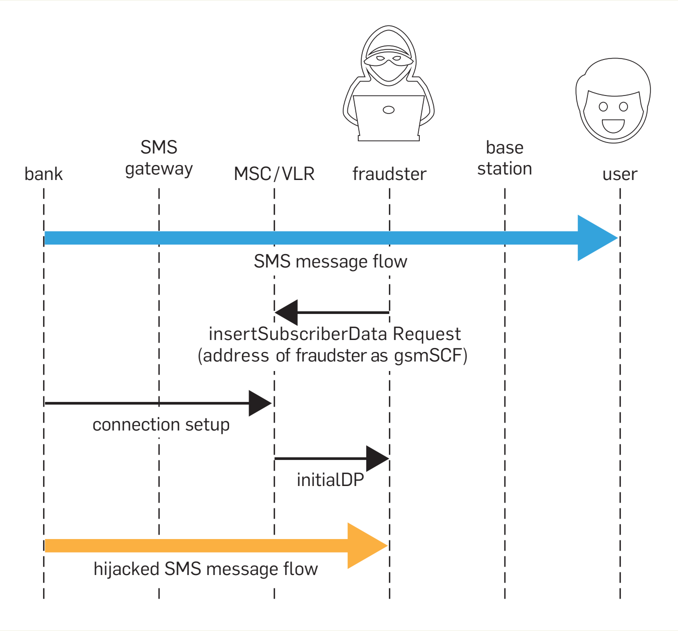
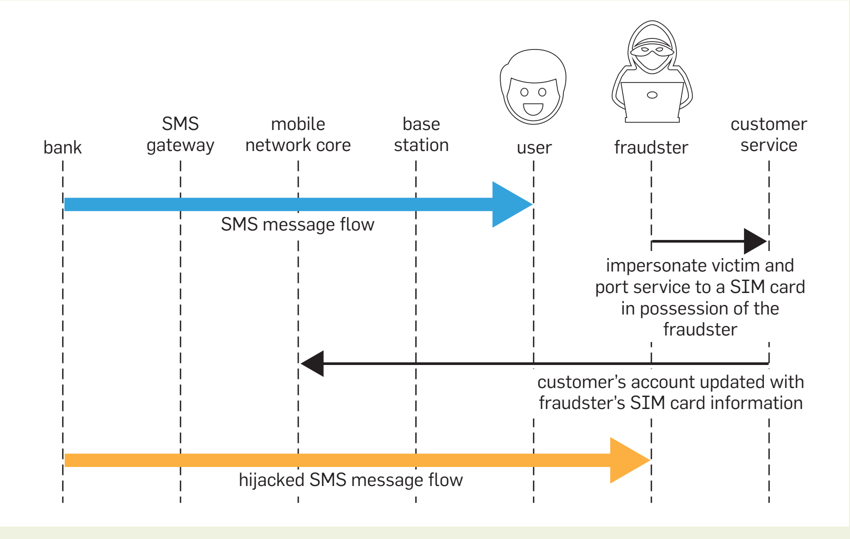


Figure 5. SIM swap attack.


bank
SMS gateway
mobile network core
base station
user
fraudster
customer service
SMS message flow
impersonate victim and port service to a SIM card in possession of the fraudster
customer's account updated with fraudster's SIM card information
hijacked SMS message flow

security risk in America and Europe, it is a much mitigated threat in African nations.[7] For example, cellular operators in Mozambique provide a means for banks to check their records for recent SIM swaps for a given account. If a SIM swap has recently occurred, the bank will deny a transaction or prevent a security token from being sent via text message.

Such a simple solution is reported to have reduced SIM swap-based banking fraud to nearly zero overnight. SIM swapping, however, is still arguably one of the biggest security risks for the average consumer of online banking and financial services. For example, the prevalence of SIM swap-driven fraud in the U.S. resulted in an official warning by New York State's Division of Consumer Protection.[21]

Final Thoughts

Despite their popularity and ease of use, SMS-based authentication tokens are arguably one of the least secure forms of two-factor authentication. This does not imply, however, that it is an invalid method for securing an online account.

True, there are a number of services that should not be used with tokens delivered via SMS—for example, banking and financial services, cryptocurrency services, and anything containing sensitive financial information, and credit card numbers. Personal email addresses also fall into this category. An email account takeover can have devastating consequences if that account is the cornerstone to the user's online digital identity.

On the other hand, there are many online services for which SMS-based tokens do suffice for the average consumer—for example, any vanilla accounts that store no sensitive or financial information, which attackers could not easily monetize, thereby discouraging them from trying to take over the account in the first place.

Other variables should be factored into the equation when deciding which multifactor authentication method is most appropriate. The security implications for a social media account for a well-known individual with millions of followers are very different from those for an account with just a handful of followers. Therefore, while using SMS as a second factor of authentication for some social media accounts is perfectly valid, it would be wise to opt for a different method for the account of a celebrity or politician.

The current security landscape is very different from that of two decades ago. Regardless of the critical nature of an online account or the individual who owns it, using a second form of authentication should always be the default option, regardless of the method chosen. In the wake of a large number of leaks and other intrusions, there are many username and password combinations out there in the wrong hands that make password spraying attacks cheap and easy to accomplish.

Related articles on queue.acm.org

VoIP: What is it good for?
Sudhir R. Ahuja and J. Robert Ensor
https://queue.acm.org/detail.cfm?id=1028897

Communications Surveillance: Privacy and Security at Risk
Whitfield Diffie and Susan Landau
https://queue.acm.org/detail.cfm?id=1613130

ACM CTO Roundtable on Mobile Devices in the Enterprise
Andrew Toy, André Charland, George Neville-Neil, Carol Realini, Steve Bourne, Mache Creeger
https://queue.acm.org/detail.cfm?id=2016038

References

1. Alfonsi, S. Hacking your phone. CBS News, 2016; https://www.cbsnews.com/video/hacking-your-phone/.
2. Cimpanu, C. Newer Diameter telephony protocol just as vulnerable as SS7. Bleeping Computer, 2018; https://www.bleepingcomputer.com/news/security/newer-diameter-telephony-protocol-just-as-vulnerable-as-ss7/.
3. Coonce, S. The most expensive lesson of my life: details of SIM port hack. Medium, 2019; https://medium.com/coinmonks/the-most-expensive-lesson-of-my-life-details-of-sim-port-hack-35de11517124.
4. Cox, J. Hackers are breaking directly into telecom companies to take over customer phone numbers. Motherboard Tech by Vice, 2020; https://www.vice.com/en_us/article/5dmbjx/how-hackers-are-breaking-into-att-tmobile-sprint-to-sim-swap-yeh.
5. Engel, T. SS7: Locate. track. manipulate. 31st Chaos Communication Congress, 2014.
6. Golde, N., Redon, K., Seifert, J.-P. 2013. Let me answer that for you: exploiting broadcast information in cellular networks. In *Proceedings of the 22nd Usenix Security Symp.* 33–48; https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_golde.pdf.
7. Greenberg, A. The SIM swap fix that the U.S. isn't using. *Wired*, 2019; https://www.wired.com/story/sim-swap-fix-carriers-banks/.
8. Honan, M. How Apple and Amazon security flaws led to my epic hacking. *Wired*, 2012; https://www.wired.com/2012/08/apple-amazon-mat-honan-hacking/.
9. Hunt, T. Beyond passwords: 2FA, U2F and Google Advanced Protection, 2018; https://www.troyhunt.com/beyond-passwords-2fa-u2f-and-google-advanced-protection/.
10. Hunt, T. Have i been pwned, 2020; https://haveibeenpwned.com/.
11. Hussain, S. R., Echeverria, M., Karim, I., Chowdhury, O., Bertino, E. 5GReasoner: a property-directed security and privacy analysis framework for 5G cellular network protocol. In *Proceedings of the ACM SIGSAC Conf. Computer and Communications Security*, 2019; 669–684; https://dl.acm.org/doi/abs/10.1145/3319535.3354263.
12. Inglesant, P.G., Sasse, M.A. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conf. Human Factors in Computing Systems*, 2010; 383–392; https://dl.acm.org/doi/10.1145/1753326.1753384.
13. Jover, R.P. LTE security and protocol exploits. *ShmooCon 2016 Proceedings*; https://shmoo.gitbook.io/2016-shmoocon-proceedings/bring_it_on/05_lte_security_and_protocol_exploits.
14. Jover, R.P. LTE security, protocol exploits and location tracking experimentation with low-cost software radio. CoRR, 2016, abs/1607.05171; https://arxiv.org/abs/1607.05171.
15. Jover, R.P. 5G protocol vulnerabilities and exploits. ShmooCon 2020; http://rogerpiquerasjover.net/5G_ShmooCon_FINAL.pdf.
16. Jover, R.P., Marojevic, V. Security and protocol exploit analysis of the 5G specifications. IEEE Access, 2019; https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8641117.
17. Kune, D.F., Koelndorfer, J., Hopper, N., Kim, Y. Location leaks on the GSM air interface. In *Proceedings of the 19th Annual Network and Distributed System Security Symp.*, 2012; https://www-users.cs.umn.edu/~hoppernj/celluloc.pdf.
18. Lee, K., Kaiser, B., Mayer, J., Narayanan, A. An empirical study of wireless carrier authentication for SIM swaps. In *Proceedings of the 16th Symp. Usable Privacy and Security*; https://www.ieee-security.org/TC/SPW2020/ConPro/papers/lee-conpro20.pdf.
19. Liu, C.-H., Chang, Y.-C., Huang, N.-F., Ling, Y.-L., Jan, H.-J. CAMEL evolution and PPS evaluation. IEEE Intelligent Network 2001 Workshop, 9–13. IEEE; https://ieeexplore.ieee.org/document/915288.
20. Mitre Corporation. Exploit SS7 to redirect phone calls/SMS. MITRE ATT&CK Framework; https://attack.mitre.org/techniques/T1449/.
21. New York State Department of Consumer Protection. ATT SIM-card switch scam; https://www.dos.ny.gov/consumerprotection/scams/att-sim.html.
22. Nohl, K. Breaking GSM phone privacy. Black Hat USA; https://srlabs.de/wp-content/uploads/2010/07/100729.Breaking.GSM_.Privacy.BlackHat1-1.pdf.
23. Nohl, K., Munaut, S. Wideband GSM sniffing. In *Proceedings of the 27th Chaos Communication Congress*; https://fahrplan.events.ccc.de/congress/2010/Fahrplan/events/4208.en.html.
24. Perez, D., Pico, J. A practical attack against GPRS/EDGE/UMTS/HSPA mobile data communications. Black Hat DC, 2011; https://media.blackhat.com/bh-dc-11/Perez-Pico/BlackHat_DC_2011_Perez-Pico_Mobile_Attacks-wp.pdf.
25. Russell, T. 2002. *Signaling System# 7*, 2 (2002). McGraw-Hill, New York, NY.
26. Shaik, A., Borgaonkar, R. New vulnerabilities in 5G networks. Black Hat 2019; https://i.blackhat.com/USA-19/Wednesday/us-19-Shaik-New-Vulnerabilities-In-5G-Networks-wp.pdf.
27. Shaik, A., Borgaonkar, R., Asokan, N., Niemi, V., Seifert, J.-P. Practical attacks against privacy and availability in 4G/LTE mobile communication systems. In *Proceedings of the 23rd Annual Network and Distributed System Security Symp*; https://www.ndss-symposium.org/wp-content/uploads/2017/09/06_5-ndss2016-slides_0.pdf.
28. Third Generation Partnership Project (3GPP) Technical Specification Group Services and System Aspects. Security architecture and procedures for 5G system. 3GPP TS 33.501, V1.0.0, 2018; https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3169.
29. Weber, J.E., Guster, D., Safonov, P., Schmidt, M.B. Weak password security: An empirical study. *Information Security J.: A Global Perspective 17*, 1 (2008), 45–54; https://dl.acm.org/doi/10.1080/10658980701824432.
30. XKCD. Password strength; https://xkcd.com/936/.

Roger Piqueras Jover is a senior security architect with the CTO Security Architecture Team at Bloomberg, where he is a technical leader in mobile security architecture and strategy, corporate network security architecture, wireless security analysis and design, and data science applied to network anomaly detection. He is a technology adviser on the security of LTE/5G mobile networks and wireless short-range networks for academia, industry, and government.

MIDDLEWARE 2021

22nd ACM/IFIP International Conference
6th – 10th December 2021
Fairmont Le Château Frontenac
Québec City, Québec, Canada

Call for Papers (1st Cycle: End of November 2020)

The annual ACM/IFIP Middleware conference is a major forum for the discussion of innovations and recent scientific advances of middleware systems with a focus on the design, implementation, deployment, and evaluation of distributed systems, platforms and architectures for computing, storage, and communication. Highlights of the conference will include a high quality single-track technical program, invited speakers, an industrial track, panel discussions involving academic and industry leaders, poster and demonstration presentations, a doctoral symposium, tutorials and workshops.

Topics of Interest: Original submissions of research papers on a diverse range of topics are sought, particularly those identifying new research directions. The topics of the interest for the conference include, but are not limited to:

- Cloud and data centers
- Virtualization, auto-scaling, provisioning, and scheduling
- Data-intensive computing (big data) and data analytics
- Stream Processing
- Middleware Systems for Machine learning
- Mobile and pervasive systems and services
- Middleware techniques for IoT, smart cities
- Fog, Edge computing
- Middleware for cyber-physical and Real-time systems
- Energy and power-aware techniques
- Event-based, publish/subscribe, and peer-to-peer solutions
- Networking, network function virtualization, software-defined networking
- Middleware for multimedia Systems
- Fault tolerance and Consistency
- Blockchains
- Middleware support for security and privacy
- Monitoring, resource management and analysis
- Middleware Design principles
- Programming abstractions and paradigms for middleware
- Reconfigurable, adaptable, and reflective approaches
- Reviews of middleware paradigms, e.g., object models, aspect orientation, etc.
- Methodologies and tools for middleware systems design, implementation, verification, and evaluation

Important Dates

First Cycle
Abstract Submission: November 20th, 2020
Full Paper Submission: December 1st, 2020
Author Notification: February 28th, 2021
Revised Submissions: April 2nd, 2021
Notifications of Revised Papers: April 22nd, 2021

Second Cycle
Abstract Submission: May 15th, 2021
Full Paper Submission: May 22nd, 2021
Author Notification: August 15th, 2021
Revised Submissions: September 15th, 2021
Notifications of Revised Papers: September 30th, 2021

General Co-Chairs

- Kaiwen Zhang, ÉTS Montréal
- Abdelouahed Gherbi, ÉTS Montréal

PC Co-Chairs

- Nalini Venkatasubramanian, University of California, Irvine
- Luís Veiga, Instituto Superior Técnico (U.Lisboa) & INESC-ID

For more details, please visit the Middleware 2021 website: http://2021.middleware-conf.org/

UNIVERSITÉ LAVAL

acm

ÉTS
Le génie pour l'industrie

DOI:10.1145/3381831

Creating efficiency in AI research will decrease its carbon footprint and increase its inclusivity as deep learning study should not require the deepest pockets.

BY ROY SCHWARTZ, JESSE DODGE, NOAH A. SMITH, AND OREN ETZIONI

Green AI

SINCE 2012, THE field of artificial intelligence (AI) has reported remarkable progress on a broad range of capabilities including object recognition, game playing, speech recognition, and machine translation.[43] Much of this progress has been achieved by increasingly large and computationally intensive deep learning models.[a] Figure 1, reproduced from Amodei et al.,[2] plots training cost increase over time for state-of-the-art deep learning models starting with AlexNet in 2012[24] to AlphaZero in 2017.[45] The chart shows an overall increase of 300,000x, with training cost doubling every few months. An even sharper trend can be observed in NLP word-embedding approaches by looking at ELMo[34] followed by BERT,[8] openGPT-2,[35] XLNet,[56] Megatron-LM,[42] T5,[36] and GPT-3.[4] An important paper[47] has estimated the carbon footprint of several NLP models and argued this trend is both environmentally unfriendly and prohibitively expensive, raising barriers to participation in NLP research. We refer to such work as Red AI.

This trend is driven by the strong focus of the AI community on obtaining "state-of-the-art" results,[b] as exemplified by the popularity of leaderboards,[53,54] which typically report accuracy (or other similar measures) but omit any mention of cost or efficiency (see, for example, leaderboards.allenai.org).[c] Despite the clear benefits of improving model accuracy, the focus on this single metric ignores the economic, environmental, and social cost of reaching the reported results.

We advocate increasing research activity in Green AI—AI research that is more environmentally friendly and inclusive. We emphasize that Red AI research has been yielding valuable scientific contributions to the field, but it has been overly dominant. We want to shift the balance toward the Green AI option—to ensure any inspired undergraduate with a laptop has the opportunity to write high-quality papers that could be accepted at premier research conferences. Specifically, we propose making efficiency a more common evaluation criterion for AI papers alongside accuracy and related measures.

[a] For brevity, we refer to AI throughout this article, but our focus is on AI research that relies on deep learning methods.

[b] Meaning, in practice, that a system's accuracy on some benchmark is greater than any previously reported system's accuracy.

[c] Some leaderboards do focus on efficiency (https://dawn.cs.stanford.edu/benchmark/).

» key insights

- The computational costs of state-of-the-art AI research has increased 300,000x in recent years. This trend, denoted Red AI, stems from the AI community's focus on accuracy while paying attention to efficiency.
- Red AI leads to a surprisingly large carbon footprint, and makes it difficult for academics, students, and researchers to engage in deep learning research.
- An alternative is Green AI, which treats efficiency as a primary evaluation criterion alonside accuracy. To measure efficiency, we suggest reporting the number of floating-point operations required to generate a result.
- Green AI research will decrease AI's environmental footprint and increase its inclusivity.

ILLUSTRATION BY LISA SHEEHAN

AI

AI research can be computationally expensive in a number of ways, but each provides opportunities for efficient improvements; for example, papers can plot performance as a function of training set size, enabling future work to compare performance even with small training budgets. Reporting the computational price tag of developing, training, and running models is a key Green AI practice (see Equation 1). In addition to providing transparency, price tags are baselines that other researchers could improve on.

Our empirical analysis in Figure 2 suggests the AI research community has paid relatively little attention to computational efficiency. In fact, as Figure 1 illustrates, the computational cost of high-budget research is increasing exponentially, at a pace that far exceeds Moore's Law.[33] Red AI is on the rise despite the well-known diminishing returns of increased cost (for example, Figure 3).

This article identifies key factors that contribute to Red AI and advocates the introduction of a simple, easy-to-compute efficiency metric that could help make some AI research greener, more inclusive, and perhaps more cognitively plausible. Green AI is part of a broader, long-standing interest in environmentally friendly scientific research (for example, see the *Journal Green Chemistry*). Computer science, in particular, has a long history of investigating sustainable and energy-efficient computing (for example, see the *Journal Sustainable Computing: Informatics and Systems*).

In this article, we analyze practices that move deep-learning research into the realm of Red AI. We then discuss our proposals for Green AI and consider related work, and directions for future research.

Figure 1. The amount of compute used to train deep learning models has increased 300,000x in six years. Figure taken from Amodei et al.[2]

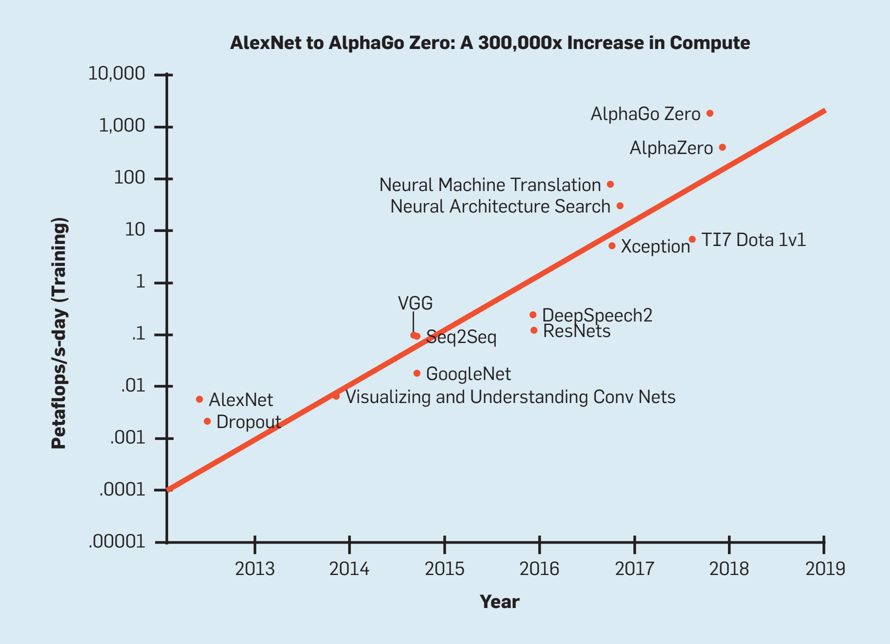
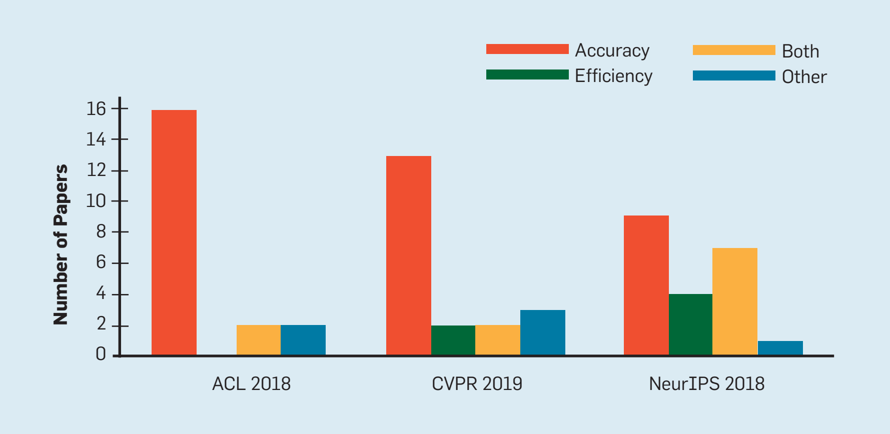


Figure 2. AI papers tend to target accuracy rather than efficiency. The figure shows the proportion of papers that target accuracy, efficiency, both or other from a random sample of 60 papers from top AI conferences.



Red AI

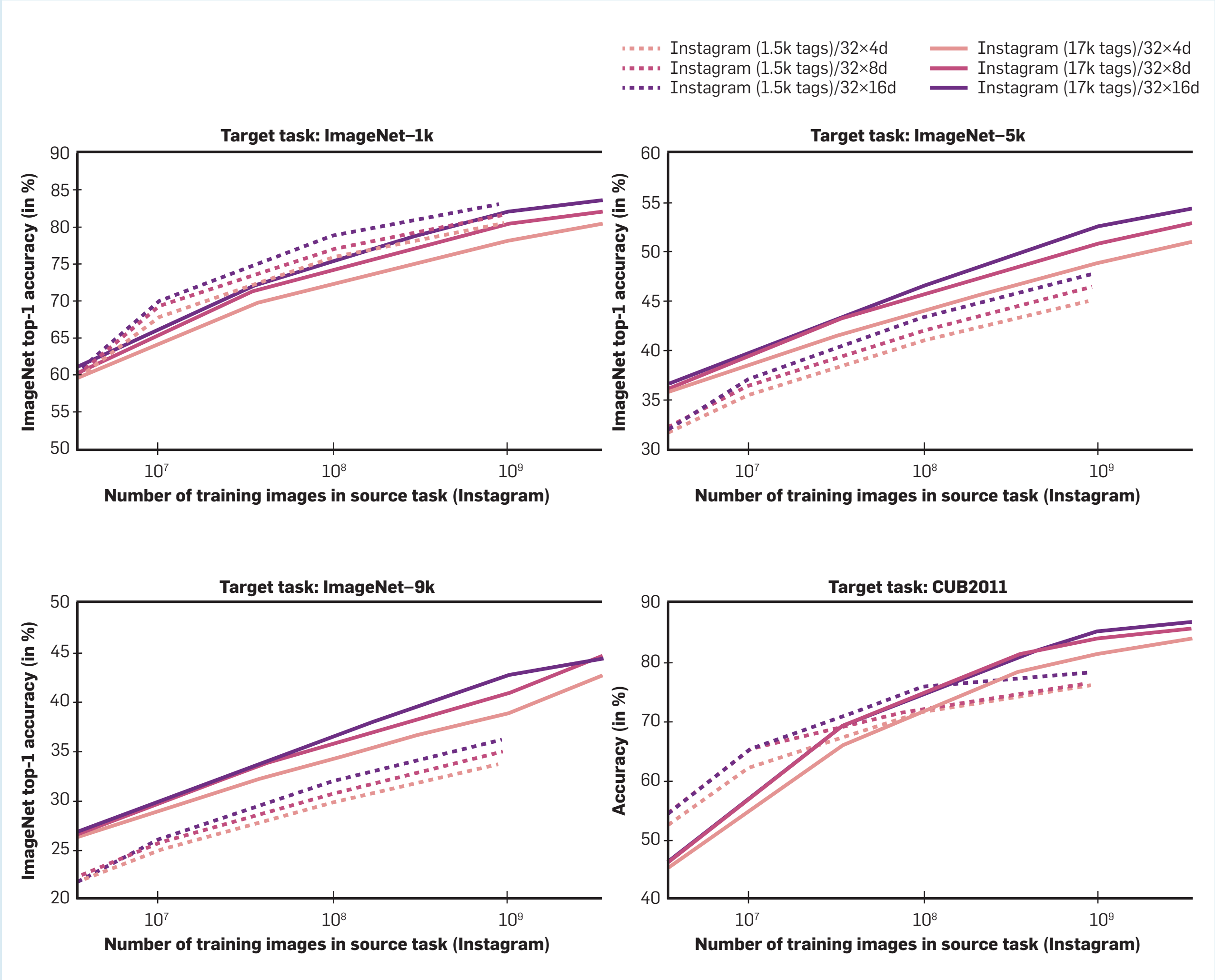
Red AI refers to AI research that seeks to improve accuracy (or related measures) through the use of massive computational power while disregarding the cost—essentially "buying" stronger results. Yet the relationship between model performance and model complexity (measured as number of parameters or inference time) has long been understood to be at best logarithmic; for a linear gain in performance, an exponentially larger model is required.[20] Similar trends exist with increasing the quantity of training data[14,48] and the number of experiments.[9,10] In each of these cases, diminishing returns come at increased computational cost.

This section analyzes the factors contributing to Red AI and shows how it is resulting in diminishing returns over time (see Figure 3). We note that Red AI work is valuable, and in fact, much of it contributes to what we know by pushing the boundaries of AI. Our exposition here is meant to highlight areas where computational expense is high, and to present each as an opportunity for developing more efficient techniques.

To demonstrate the prevalence of Red AI, we randomly sampled 60 papers from top AI conferences (ACL, NeurIPS, and CVPR).[d] For each paper we noted whether the authors claim their main contribution to be (a) an improvement to accuracy or some related measure, (b) an improvement to efficiency, (c) both, or (d) other. As shown in Figure 2, in all conferences we considered, a large majority of the papers target accuracy (90% of ACL papers, 80% of NeurIPS papers and 75% of CVPR papers). Moreover, for both empirical AI conferences (ACL

[d] https://acl2018.org; https://nips.cc/Conferences/2018; and http://cvpr2019.thecvf.com.

Figure 3. Diminishing returns of training on more data: object detection accuracy increases linearly as the number of training examples increases exponentially.[30]



and CVPR) only a small portion (10% and 20% respectively) argue for a new efficiency result.[e] This highlights the focus of the AI community on measures of performance such as accuracy, at the expense of measures of efficiency such as speed or model size. In this article, we argue that a larger weight should be given to the latter.

To better understand the different ways in which AI research can be red, consider an AI result reported in a scientific paper. This result typically characterizes a model trained on a training dataset and evaluated on a test dataset, and the process of developing that model often involves multiple experiments to tune its hyperparameters. We thus consider three dimensions that capture much of the computational cost of obtaining such a result: the cost of executing the model on a single (*E*)xample (either during training or at inference time); the size of the training (*D*)ataset, which controls the number of times the model is executed during training, and the number of (*H*)yperparameter experiments, which controls how many times the model is trained during model development. The total cost of producing a (*R*)esult in machine learning increases linearly with each of these quantities. This cost can be estimated as follows:

$$Cost\,(R) \propto E \cdot D \cdot H$$

Equation 1. The equation of Red AI: The cost of an AI (*R*)esult grows linearly with the cost of processing a single (*E*)xample, the size of the training (*D*)ataset and the number of (*H*)yperparameter experiments.

Equation 1 is a simplification (for example, different hyperparameter assignments can lead to different costs for processing a single example). It also ignores other factors such as the number of training epochs or data augmentation. Nonetheless, it illustrates three quantities that are each an important factor in the total cost of generating a result. Next, we consider each quantity separately.

Expensive processing of one example. Our focus is on neural models, where it

[e] Interestingly, many NeurIPS papers included convergence rates or regret bounds that describe performance as a function of examples or iterations, thus targeting efficiency (55%). This indicates an increased awareness of the importance of this concept, at least in theoretical analyses.

is common for each training step to require inference, so we discuss training and inference cost together as "processing" an example (though see discussion below). Some works have used increasingly large models in terms of, for example, model parameters, and as a result, in these models, performing inference can require a lot of computation, and training even more so. For instance, Google's BERT-large[8] contains roughly 350 million parameters. OpenAI's openGPT2-XL model[35] contains 1.5 billion parameters. AI2, our home organization, released Grover,[57] also containing 1.5 billion parameters. NVIDIA released Megatron-LM,[42] containing over 8 billion parameters. Google's T5-11B[36] contains 11 billion parameters. Most recently, openAI released openGPT-3,[4] containing 175 billion parameters. In the computer vision community, a similar trend is observed (Figure 1).

Such large models have high costs for processing each example, which leads to large training costs. BERT-large was trained on 64 TPU chips for four days at an estimated cost of $7,000. Grover was trained on 256 TPU chips for two weeks, at an estimated cost of $25,000. XLNet had a similar architecture to BERT-large, but used a more expensive objective function (in addition to an order of magnitude more data), and was trained on 512 TPU chips for 2.5 days, costing more than $60,000.[f] It is impossible to reproduce the best BERT-large results or XLNet results using a single GPU,[g] and models such as openGPT2 are too large to be used in production.[h] Specialized models can have even more extreme costs, such as AlphaGo, the best version of which required 1,920 CPUs and 280 GPUs to play a single game of Go,[44] with an estimated cost to reproduce this experiment of $35,000,000.[i,j]

When examining variants of a single model (for example, BERT-small and BERT-large) we see that larger models can have stronger performance, which is a valuable scientific contribution. However, this implies the financial and environmental cost of increasingly large AI models will not decrease soon, as the pace of model growth far exceeds the resulting increase in model performance.[18] As a result, more and more resources are going to be required to keep improving AI models by simply making them larger.

Finally, we note that in some cases the price of processing one example might be different at training and test time. For instance, some methods target efficient inference by learning a smaller model based on the large trained model. These models often do not lead to more efficient training, as the cost of E is only reduced at inference time. Models used in production typically have computational costs dominated by inference rather than training, but in research training is typically much more frequent, so we advocate studying methods for efficient processing of one example in both training and inference.

Processing many examples. Increased amounts of training data have also contributed to progress in state-of-the-art performance in AI. BERT-large had top performance in 2018 across many NLP tasks after training on three billion word-pieces. XLNet outperformed BERT after training on 32 billion word-pieces, including part of Common Crawl; openGPT-2-XL trained on 40 billion words; FAIR's RoBERTa[28] was trained on 160GB of text, roughly 40 billion word-pieces, requiring around 25,000 GPU hours to train. T5-11B[36] was trained on 1 trillion tokens, 300 times more than BERT-large. In computer vision, researchers from Facebook[30] pretrained an image classification model on 3.5 billion images from Instagram, three orders of magnitude larger than existing labeled image datasets such as Open Images.[k]

The use of massive data creates barriers for many researchers to reproducing the results of these models, and to training their own models on the same setup (especially as training for multiple epochs is standard). For example, the July 2019 Common Crawl contains 242TB of uncompressed data,[l] so even storing the data is expensive. Finally, as in the case of model size, relying on more data to improve performance is notoriously expensive because of the diminishing returns of adding more data.[48] For instance, Figure 3, taken from Mahajan et al.,[30] shows a logarithmic relation between the object recognition top-1 accuracy and the number of training examples.

Massive number of experiments. Some projects have poured large amounts of computation into tuning hyperparameters or searching over neural architectures, well beyond the reach of most researchers. For instance, researchers from Google[59] trained over 12,800 neural networks in their neural architecture search to improve performance on object detection and language modeling. With a fixed architecture, researchers from DeepMind[31] evaluated 1,500 hyperparameter assignments to demonstrate that an LSTM language model[17] can reach state-of-the-art perplexity results. Despite the value of this result in showing that the performance of an LSTM does not plateau after only a few hyperparameter trials, fully exploring the potential of other competitive models for a fair comparison is prohibitively expensive.

The value of massively increasing the number of experiments is not as well studied as the first two discussed previously. In fact, the number of experiments performed during model construction is often underreported. Nonetheless, evidence for a logarithmic relation exists here as well.[9,10]

Discussion. The increasing costs of AI experiments offer a natural economic motivation for developing more efficient AI methods. It might be the case that at a certain point prices will be too high, forcing even researchers with large budgets to develop more efficient methods. Our analysis in Figure 2 shows that currently most effort is still being dedicated to accuracy rather than efficiency. At the same time, AI technology is already very expensive to train or execute, which limits the ability of many researchers to study it, and of practitioners to adopt it. Combined with environmental pricetag of AI,[47] we believe more effort should be devoted toward efficient AI solutions.

[f] https://syncedreview.com/2019/06/27/the-staggering-cost-of-training-sota-aimodels/

[g] See https://github.com/google-research/bert and https://github.com/zihangdai/xlnet.

[h] https://towardsdatascience.com/too-big-to-deploy-how-gpt-2-is-breakingproduction-63ab29f0897c

[i] https://www.yuzeh.com/data/agz-cost.html

[j] Recent versions of AlphaGo are far more efficient.[46]

[k] https://opensource.google.com/projects/open-images-dataset

[l] http://commoncrawl.org/2019/07/

We want to reiterate that Red AI work is extremely valuable, and in fact, much of it contributes to what we know about pushing the boundaries of AI. Indeed, there is value in pushing the limits of model size, dataset size, and the hyperparameter search budget.

In addition, Red AI can provide opportunities for future work to promote efficiency; for example, evaluating a model on varying amounts of training data will provide an opportunity for future researchers to build on the work without needing a budget large enough to train on a massive dataset. Currently, despite the massive amount of resources put into recent AI models, such investment still pays off in terms of downstream performance (albeit at an increasingly lower rate). Finding the point of saturation (if such exists) is an important question for the future of AI. Moreover, Red AI costs can even sometimes be amortized, because a Red AI trained module may be reused by many research projects as a built-in component, which doesn't require retraining.

The goal of this article is twofold: first, we want to raise awareness to the cost of Red AI and encourage researchers that use such methods to take steps to allow for more equitable comparisons, such as reporting training curves. Second, we want to encourage the AI community to recognize the value of work by researchers that take a different path, optimizing efficiency rather than accuracy. Next, we turn to discuss concrete measures for making AI more green.

GREEN AI

The term Green AI refers to AI research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent. Whereas Red AI has resulted in rapidly escalating computational (and thus carbon) costs, Green AI promotes approaches that have favorable performance/efficiency trade-offs. If measures of efficiency are widely accepted as important evaluation metrics for research alongside accuracy, then researchers will have the option of focusing on the efficiency of their models with positive impact on both inclusiveness and the environment. Here, we review several measures of efficiency that could be reported and optimized, and advocate one particular measure—FPO—which we argue should be reported when AI research findings are published.

Some projects have poured large amounts of computation into tuning hyperparameters or searching over neural architectures, well beyond the reach of most researchers.

Measures of efficiency. To measure efficiency, we suggest reporting the amount of work required to generate a result. Specifically, the amount of work required to train a model, and if applicable, the aggregated amount of work required for all hyperparameter tuning experiments. As the cost of an experiment decomposes into the cost of a processing a single example, the size of the dataset, and the number of experiments (Equation 1), reducing the amount of work in each of these steps will result in AI that is more green.

We do encourage AI practitioners to use efficient hardware to reduce energy costs, but the dramatic increase in computational cost observed over recent years is primarily from modeling and algorithmic choices; our focus is on how to incorporate efficiency there. When reporting the amount of work done by a model, we want to measure a quantity that allows for a fair comparison between different models. As a result, this measure should ideally be stable across different labs, at different times, and using different hardware.

Carbon emission. Carbon emission is appealing as it is a quantity we want to directly minimize. Nonetheless it is difficult to measure the exact amount of carbon released by training or executing a model, and accordingly—generating an AI result, as this amount depends highly on the local electricity infrastructure (though see initial efforts by Henderson et al.[16] and Lacoste et al.[25]). As a result, it is not comparable between researchers in different locations or even the same location at different times.[16]

Electricity usage. Electricity usage is correlated with carbon emission while being time- and location-agnostic. Moreover, GPUs often report the amount of electricity each of their cores consume at each time point, which facilitates the estimation of the total amount of electricity consumed by generating an AI result. Nonetheless, this measure is hardware dependent, and as a result does not allow for a fair comparison between different models developed on different machines.

Elapsed real time. The total running time for generating an AI result is a natural measure for efficiency, as all other things being equal, a faster model is doing less computational work. Nonetheless, this measure is highly influenced by factors such as the underlying hardware, other jobs running on the same machine, and the number of cores used. These factors hinder the comparison between different models, as well as the decoupling of modeling contributions from hardware improvements.

Number of parameters. Another common measure of efficiency is the number of parameters (learnable or total) used by the model. As with runtime, this measure is correlated with the amount of work. Unlike the other measures described previously, it does not depend on the underlying hardware. Moreover, this measure also highly correlates with the amount of memory consumed by the model. Nonetheless, different algorithms make different use of their parameters, for instance by making the model deeper vs. wider. As a result, different models with a similar number of parameters often perform different amounts of work.

FPO. As a concrete measure, we suggest reporting the total number of floating-point operations (FPO) required to generate a result.[m] FPO provides an estimate of the amount of work performed by a computational process. It is computed analytically by defining a cost to two base operations, ADD and MUL. Based on these operations, the FPO cost of any machine learning abstract operation (for example, a tanh operation, a matrix multiplication, a convolution operation, or the BERT model) can be computed as a recursive function of these two operations. FPO has been used in the past to quantify the energy footprint of a model,[13,32,50,51] but is not widely adopted in AI. FPO has several appealing properties. First, it directly computes the amount of work done by the running machine when executing a specific instance of a model and is thus tied to the amount of energy consumed. Second, FPO is agnostic to the hardware on which the model is run. This facilitates fair comparisons between different approaches, unlike the measures described above. Third, FPO is often correlated with the running time of the model[5] (though see discussion below). Unlike asymptotic runtime, FPO also considers the amount of work done at each time step.

The term Green AI refers to AI research that yields novel results while taking into account the computational cost, encouraging a reduction in resources spent.

Several packages exist for computing FPO in various neural network libraries,[n] though none of them contains all the building blocks required to construct all modern AI models. We encourage the builders of neural network libraries to implement such functionality directly.

Discussion. Efficient machine learning approaches have received attention in the research community but are generally not motivated by being green. For example, a significant amount of work in the computer vision community has addressed efficient inference,[13,38,58] which is necessary for real-time processing of images for applications like self-driving cars,[27,29,37] or for placing models on devices such as mobile phones.[18,40] Most of these approaches only minimize the cost of processing a single example, while ignoring the other two red practices discussed perviously.[o] Other methods to improve efficiency aim to develop more efficient architectures, starting from the adoption of graphical processing units (GPU) to AI algorithms, which was the driving force behind the deep learning revolution, up to more recent development of hardware such as tensor processing units (TPUs[22]).

The examples here indicate the path to making AI green depends on how it is used. When developing a new model, much of the research process involves training many model variants on a training set and performing inference on a small development set. In such a setting, more efficient training procedures can lead to greater savings, while in a production setting more efficient inference can be more important. We advocate for a holistic view of computational savings which doesn't sacrifice in some areas to make advances in others.

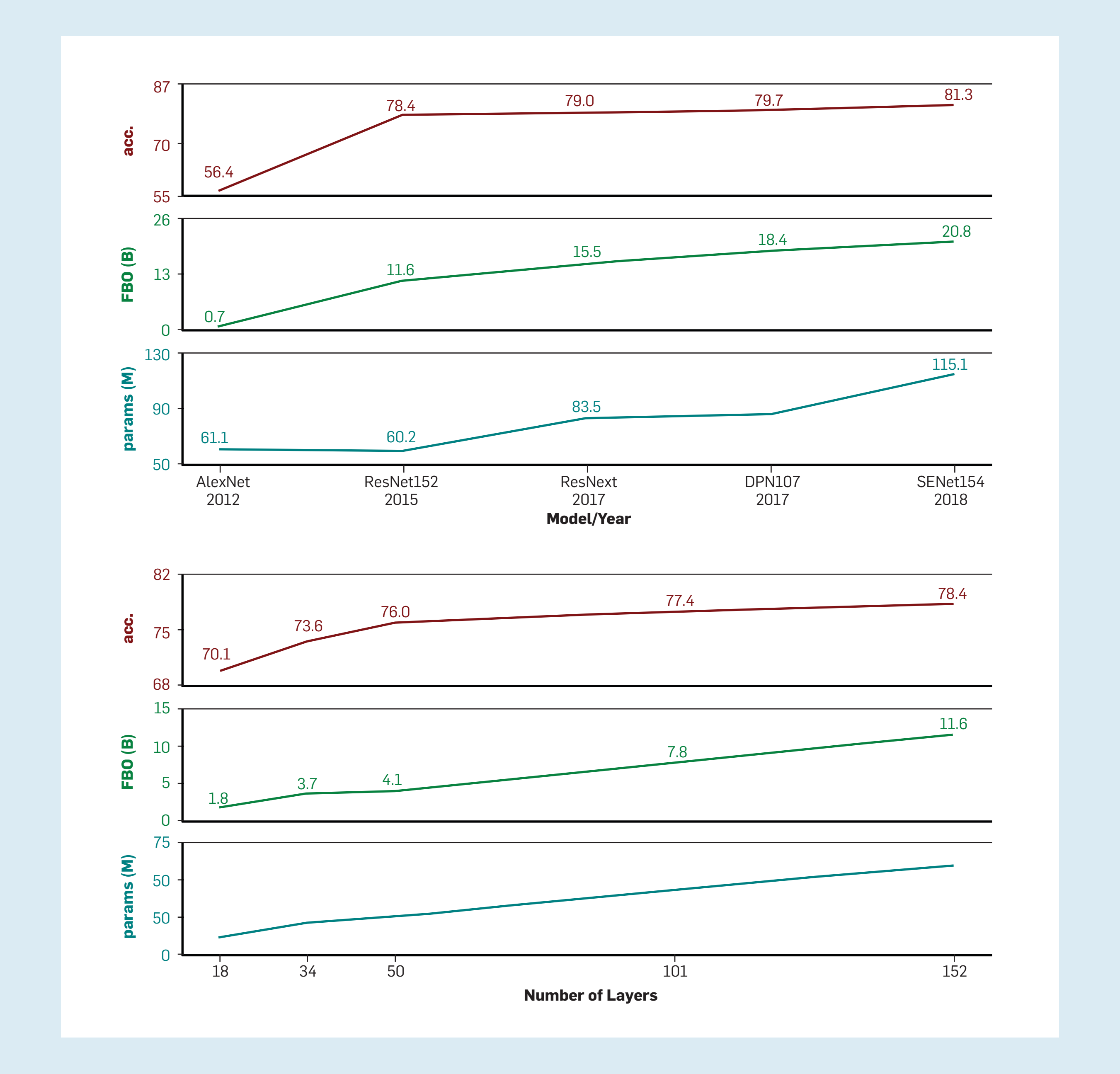
FPO has some limitations. Most importantly, the energy consumption of a

[m] Floating point operations are often referred to as FLOP(s), though this term is not uniquely defined.[13] To avoid confusion, we use the term FPO.

[n] For example, https://github.com/Swall0w/torchstat; https://github.com/Lyken17/pytorch-OpCounter

[o] In fact, creating smaller models often results in longer running time, so mitigating the different trends might be at odds.[52]

Figure 4. Increase in FPO leads to diminishing return for object detection top-1 accuracy. Plots (bottom to top): model parameters (in million), FPO (in billions), top-1 accuracy on ImageNet. 4(a). Leading object recognition models: AlexNet,[24] ResNet,[15] ResNext,[55] DPN107,[6] SENet154.[19] 4(b): Comparison of different sizes (measured by the number of layers) of the ResNet model.[15]



model is not only influenced by the amount of work, but also from other factors such as the communication between the different components, which is not captured by FPO. As a result, FPO doesn't always correlate with other measures such as runtime[21] and energy consumption.[16] Second, FPO targets the number of operations performed by a model, while ignoring other potential limiting factors for researchers such as the memory used by the model, which can often lead to additional energy and monetary costs.[29] Finally, the amount of work done by a model largely depends on the model implementation, as two different implementations of the same model could result in very different amounts of processing work. Due to the focus on the modeling contribution, the AI community has traditionally ignored the quality or efficiency of models' implementation.[p] We argue the time to reverse this norm has come, and that exceptionally good implementations that lead to efficient models should be credited by the AI community.

FPO cost of existing models. To demonstrate the importance of reporting the amount of work, we present FPO costs for several existing models.[q] Figure 4(a) shows the number of parameters and FPO of several leading object recognition models, as well as their performance on the ImageNet

[p] We consider this exclusive focus on the final prediction another symptom of Red AI.

[q] These numbers represent FPO per inference, that is, the work required to process a single example.

dataset.[7,r] A few trends are observable. First, as discussed earlier, models get more expensive with time, but the increase in FPO does not lead to similar performance gains. For instance, an increase of almost 35% in FPO between ResNet and ResNext (second and third points in graph) resulted in a 0.5% top-1 accuracy improvement. Similar patterns are observed when considering the effect of other increases in model work. Second, the number of model parameters does not tell the whole story: AlexNet (first point in the graph) actually has more parameters than ResNet (second point), but dramatically less FPO, and also much lower accuracy.

Figure 4(b) shows the same analysis for a single object recognition model, ResNet,[15] while comparing different versions of the model with different numbers of layers. This creates a controlled comparison between the different models, as they are identical in architecture, except for their size (and accordingly, their FPO cost). Once again, we notice the same trend: the large increase in FPO cost does not translate to a large increase in performance.

Additional ways to promote Green AI. There are many ways to encourage research that is more green. In addition to reporting the FPO cost for each term in Equation 1, we encourage researchers to report budget/performance curves where possible. For example, training curves provide opportunities for future researchers to compare at a range of different budgets and running experiments with different model sizes provides valuable insight into how model size impacts performance. In a recent paper,[9] we observed that the claim as to which model performs best depends on the computational budget available during model development. We introduced a method for computing the expected best validation performance of a model as a function of the given budget. We argue that reporting this curve will allow users to make wiser decisions about their selection of models and highlight the stability of different approaches.

We further advocate for making efficiency an official contribution in major AI conferences by advising reviewers to recognize and value contributions that do not strictly improve state of the art but have other benefits such as efficiency. Finally, we note that the trend of releasing pretrained models publicly is a green success, and we would like to encourage organizations to continue to release their models in order to save others the costs of retraining them.

Related Work

Recent work has analyzed the carbon emissions of training deep NLP models[47] and concluded that computationally expensive experiments can have a large environmental and economic impact. With modern experiments using such large budgets, many researchers (especially those in academia) lack the resources to work in many high-profile areas; increased value placed on computationally efficient approaches will allow research contributions from more diverse groups. We emphasize that the conclusions of Stubell et al.[47] are the result of long-term trends, and are not isolated within NLP, but hold true across machine learning.

While some companies offset electricity usage by purchasing carbon credits, it is not clear that buying credits is as effective as using less energy. In addition, purchasing carbon credits is voluntary; Google cloud[s] and Microsoft Azure[t] purchase carbon credits to offset their spent energy, but Amazon's AWS[u] (the largest cloud computing platform[v]) only covered 50% of its power usage with renewable energy.

The push to improve state-of-the-art performance has focused the research community's attention on reporting the single best result after running many experiments for model development and hyperparameter tuning. Failure to fully report these experiments prevents future researchers from understanding how much effort is required to reproduce a result or extend it.[9]

Our focus is on improving efficiency in the machine learning community, but machine learning can also be used as a tool for work in areas like climate change. For example, machine learning has been used for reducing emissions of cement plants[1] and tracking animal conservation outcomes,[12] and is predicted to be useful for forest fire management.[39] Undoubtedly these are important applications of machine learning; we recognize they are orthogonal to the content of this article.

Conclusion

The vision of Green AI raises many exciting research directions that help to overcome the challenges of Red AI. Progress will find more efficient ways to allocate a given budget to improve performance, or to reduce the computational expense with a minimal reduction in performance. Also, it would seem that Green AI could be moving us in a more cognitively plausible direction as the brain is highly efficient.

It is important to reiterate that we see Green AI as a valuable option, not an exclusive mandate—of course, both Green AI and Red AI have contributions to make. Our goals are to augment Red AI with green ideas, like using more efficient training methods, and reporting training curves; and to increase the prevalence of Green AI by highlighting its benefits, advocating a standard measure of efficiency. Here, we point to a few important green research directions, and highlight a few open questions.

Research on building space- or time-efficient models is often motivated by fitting a model on a small device (such as a phone) or fast enough to process examples in real time, such as image captioning for the blind (as discussed previously). Here, we argue for a far broader approach that promotes efficiency for all parts of the AI development cycle.

Data efficiency has received significant attention over the years.[23,41,49] Modern research in vision and NLP often involves first pretraining a model on large "raw" (unannotated) data then finetuning it to a task of interest through supervised learning. A strong result in this area often involves achieving similar performance to a baseline with fewer training examples or fewer gradient steps. Most recent work has addressed fine-tuning data,[34] but pretraining efficiency is also important. In either case, one simple technique to improve in this area is to

[r] Numbers taken from https://github.com/sovrasov/flops-counter.pytorch.

[s] https://cloud.google.com/sustainability/

[t] https://www.microsoft.com/en-us/environment/carbon

[u] https://aws.amazon.com/about-aws/sustainability/

[v] https://tinyurl.com/y2kob969

simply report performance with different amounts of training data. For example, reporting performance of contextual embedding models trained on 10 million, 100 million, 1 billion, and 10 billion tokens would facilitate faster development of new models, as they can first be compared at the smallest data sizes.

Research here is of value not just to make training less expensive, but because in areas such as low resource languages or historical domains it is extremely difficult to generate more data, so to progress we must make more efficient use of what is available.

Finally, the total number of experiments run to get a final result is often underreported and underdiscussed.[9] The few instances researchers have of full reporting of the hyperparameter search, architecture evaluations, and ablations that went into a reported experimental result has surprised the community.[47] While many hyperparameter optimization algorithms exist, which can reduce the computational expense required to reach a given level of performance,[3,11] simple improvements here can have a large impact. For example, stopping training early for models that are clearly underperforming can lead to great savings.[26]

Acknowledgment. This research was conducted at the Allen Institute for AI. C

References

1. Acharyya, P., Rosario, S.D., Flor, F., Joshi, R., Li, D., Linares, R, and Zhang, H. Autopilot of cement plants for reduction of fuel consumption and emissions. In *Proceedings of ICML Workshop on Climate Change*, 2019.
2. Amodei, D. and Hernandez, D. AI and compute, 2018. Blog post.
3. Bergstra, J.S., Bardenet, R., Bengio, Y. and Kégl, B. Algorithms for hyper-parameter optimization. In *Proceedings of NeurIPS*, 2011.
4. Brown, T.B. et al. Language models are few-shot learners, 2020; arXiv:2005.14165.
5. Canziani, A., Paszke, A. and Culurciello, E. An analysis of deep neural network models for practical applications. In *Proceedings of ISCAS*, 2017.
6. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S. and Feng, J. Dual path networks. In *Proceedings of NeurIPS*, 2017.
7. Deng, J., Dong, W., Socher, R., Li, L-J, Li, K. and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, 2009.
8. Devlin, J., Chang, M.W., Lee, K., and Toutanova, K. BERT: Pretraining of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, 2019.
9. Dodge, J., Gururangan, S., Card, D., Schwartz, R. and Smith, N.A. Show your work: Improved reporting of experimental results. In *Proceedings of EMNLP*, 2019.
10. Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H. and Smith, N.A. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020; arXiv:2002.06305.
11. Dodge, J., Jamieson, K. and Smith, N.A. Open loop hyperparameter optimization and determinantal point processes. In *Proceedings of AutoML*, 2017.
12. Duhart, C., Dublon, G., Mayton, B., Davenport, G. and Paradiso, J.A. Deep learning for wildlife conservation and restoration efforts. In *Proceedings of ICML Workshop on Climate Change*, 2019.
13. Gordon, A., Eban, E., Nachum, O., Chen, B., Wu, H., Yang, T-J, and Choi, E. MorphNet: Fast & simple resource-constrained structure learning of deep networks. In *Proceedings of CVPR*, 2018.
14. Halevy, A., Norvig, P. and Pereira, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems 24* (2009), 8–12.
15. He, K., Zhang, X., Ren, S. and Sun, J. Deep residual learning for image recognition. In *Proceedings of CVPR*, 2016.
16. Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning, 2020; arXiv:2002.05651.
17. Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation 9*, 8 (1997), 1735–1780.
18. Howard, A.G. et al. MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017; arXiv:1704.04861.
19. Hu, J., Shen, L. and Sun, G. Squeeze-and-excitation networks. In *Proceedings of CVPR*, 2018.
20. Huang, J. et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of CVPR*, 2017.
21. Jeon, Y. and Kim, J. Constructing fast network through deconstruction of convolution. In *Proceedings of NeurIPS*, 2018.
22. Jouppi, N.P. et al. In-datacenter performance analysis of a tensor processing unit. In *Proceedings of ISCA 1*, 1 (2017), Publ. date: June 2020.
23. Kamthe, S. and Deisenroth, M.P. Data-efficient reinforcement learning with probabilistic model predictive control. In *Proceedings of AISTATS*, 2018.
24. Krizhevsky, A., Sutskever, I. and Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Proceedings of NeurIPS*, 2012.
25. Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T. Quantifying the carbon emissions of machine learning. In *Proceedings of the Climate Change AI Workshop*, 2019.
26. Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A. and Talwalkar, A. Hyperband: Bandit-based configuration evaluation for hyperparameter optimization. In *Proceedings of ICLR*, 2017.
27. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. Fu, C-Y and Berg, A.C. SSD: Single shot multibox detector. In *Proceedings of ECCV*, 2016.
28. Liu, Y. et al. RoBERTa: A robustly optimized BERT pretraining approach, 2019; arXiv:1907.11692.
29. Ma, N., Zhang, X., Zheng, H.T and Sun, J. ShuffleNet V2: Practical guidelines for efficient cnn architecture design. In *Proceedings of ECCV*, 2018.
30. Mahajan, D. et al. Exploring the limits of weakly supervised pretraining, 2018; arXiv:1805.00932.
31. Melis, G., Dyer, C. and Blunsom, P. On the state of the art of evaluation in neural language models. In *Proceedings of EMNLP*, 2018.
32. Molchanov, P., Tyree, S., Karras, T., Aila, T. and Kautz, J. Pruning convolutional neural networks for resource efficient inference. In *Proceedings of ICLR*, 2017.
33. Moore, G.E. Cramming more components onto integrated circuits, 1965.
34. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of NAACL*, 2018.
35. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. Language m odels are unsupervised multitask learners.. OpenAI Blog, 2019.
36. Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer, 2019; arXiv:1910.10683.
37. Rastegari, M., Ordonez, V., Redmon, J. and Farhadi, A. Xnornet: Imagenet classification using binary convolutional neural networks. In *Proceedings of ECCV*, 2016.
38. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of CVPR*, 2016.
39. Rolnick, D. et al. Tackling climate change with machine learning, 2019; arXiv:1905.12616.
40. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of CVPR*, 2018.
41. Schwartz, R., Thomson, S. and Smith, N.A. SoPa: Bridging CNNs, RNNs, and weighted finite-state machines. In *Proceedings of ACL*, 2018.
42. Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., Catanzaro, B. Megatron-LM: Training multi-billion parameter language models using GPU model parallelism, 2019; arXiv:1909.08053.
43. Shoham, Y. et al. The AI index 2018 annual report. AI Index Steering Committee, Human-Centered AI Initiative, Stanford University; http:// cdn.aiindex.org/2018/AI%20Index%202018%20Annual%20Report.pdf.
44. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature 529*, 7587 (2016) 484.
45. Silver, D. et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017; arXiv:1712.01815.
46. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature 550*, 7676 (2017), 354.
47. Strubell, E., Ganesh, A. and McCallum, A. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL*, 2019.
48. Sun, C., Shrivastava, A., Singh, S. and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of ICCV, 2017.
49. Tsang, I., Kwok, J.T. and Cheung, P.M. Core vector machines: Fast SVM training on very large data sets. *JMLR 6* (Apr. 2005), 363–392.
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. Attention is all you need. In *Proceedings of NeurIPS*, 2017.
51. Veniat, T. and Denoyer, L. Learning time/memory-efficient deep architectures with budgeted super networks. In *Proceedings of CVPR*, 2018.
52. Walsman, A., Bisk, Y., Gabriel, S., Misra, D., Artzi, Y., Choi, Y. and Fox, D. Early fusion for goal directed robotic vision. In *Proceedings of IROS*, 2019.
53. Wang, A. Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. SuperGLUE: A stickier benchmark for general-purpose language understanding systems, 2019; arXiv:1905.00537.
54. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of ICLR*, 2019.
55. Xie, S., Girshick, R., Dollar, P., Tu, Z. and He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of CVPR*, 2017.
56. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q.V. XLNet: Generalized autoregressive pretraining for language understanding, 2019; arXiv:1906.08237.
57. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. and Choi, Y. Defending against neural fake news, 2019; arXiv:1905.12616.
58. Zhang, X., Zhou, X., Lin, M. and Sun, J. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of CVPR*, 2018.
59. Zoph, B. and Le, Q.V. Neural architecture search with reinforcement learning. In *Proceedings of ICLR*, 2017.

Roy Schwartz (roys@allenai.org) is Senior Lecturer at the Hebrew University of Jerusalem, Israel.

Jesse Dodge (dodgejesse@gmail.com), Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA.

Noah A. Smith (noah@allenai.org) is a professor of computer science and engineering at the University of Washington and senior research manager for the AllenNLP team at Allen Institute for AI and, Seattle, WA, USA.

Oren Etzioni (orene@allenai.org) is Chief Executive Officer of the Allen Institute for AI, and a professor of computer science at the University of Washington, Seattle, WA, USA.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/green-ai

DOI:10.1145/3408864

Tracing the relationship between pathological personality traits and insider cyber sabotage.

BY MICHELE MAASBERG, CRAIG VAN SLYKE, SELWYN ELLIS, AND NICOLE BEEBE

The Dark Triad and Insider Threats in Cyber Security

"I WAS DISMAYED to learn this weekend about a Tesla employee who had conducted quite extensive and damaging sabotage to our operations. This included making direct code changes to the Tesla Manufacturing Operating System under false usernames and exporting large amounts of highly sensitive Tesla data to unknown third parties."

—Tesla CEO Elon Musk in an email to Tesla employees.[6]

Insider cyber sabotage[4] such as that mentioned by Mr. Musk is one of the reasons cyber security remains a top managerial concern. Insider threats, such as the Tesla sabotage, are among the greatest of these security concerns.[24] A major reason for this is that insider security breaches are seen as more costly than those from outsiders.[16,20]

Understanding the individual, social, and organizational influences on insider threats is important to the development of security-related policies and controls. Cyber sabotage as part of a broader insider threat issue is addressed in the context of an organizational security risk management plan. Such plans should include security controls intended to mitigate the risk of a human threat from the inside. In the U.S., in some cases in which classified material is involved, formal insider threat cyber security programs are mandated by Presidential Executive Order.

The security controls prescribed by insider threat programs often include automated employee monitoring systems for detection, education and training programs for awareness.[9] These controls often include technical and behavioral indicators derived from the observed psychological traits and specific behaviors of high-risk insiders. These indicators should be based on empirical evidence in order to avoid false accusations that harm employees[7] and negative ethical and legal consequences associated with biased systems.[12]

Insiders possess unique personal predispositions, stressors, and concerning behaviors that have been identified as risk factors; these have been included in models of insider threat behaviors.[8,18] Past research suggests that robust cybersecurity systems include psychological or personality factors in their design.[9] Several insider threat frameworks include personal predispositions (including personality traits) as the origin point of threat behaviors.[8,17,18] This suggests it is important to recognize personal factors, especially personality traits, before

» key insights

- Malicious insider threats often exhibit the malevolent personality traits subclinical narcissism, psychopathy, and Machiavellianism known as the dark triad.
- The cost of hiring a toxic worker typically exceeds any benefit that they might bring to an organization.

IMAGE BY ALICIA KUBISTA/ANDRIJ BORYS ASSOCIATES

NARCISSISM
PSYCHOPATHY
THE
DARK
TRIAD
MACHIAVELLIANISM

they lead to malicious behaviors. Such recognition can be the earliest point of threat agent identification.

Much of the existing research into personality traits and cybersecurity is based on case studies, anecdotal evidence, or conceptual reasoning. There is a lack of quantitative empirical evidence to guide our understanding of the relationship between personality traits and insider threats.[13] Understanding the role of traits related to antisocial behavior in malicious insider threats is especially important due to the link between these traits and malevolent behavior. The findings of our research may help enhance and extend existing models and frameworks including advanced technical systems.

In this article, we focus on a set of pathological personality traits known as the dark triad. Evidence from recent insider threat cases leads us to believe these traits may correlate with intentions to engage in malicious behavior.[23] After discussing insider threats and the dark triad traits, we present results from an empirical study that illustrate the relationship between the dark triad traits and malicious intent. We then discuss the importance of these results and make recommendations for security managers and practitioners based on our findings. Despite the inclusion of personality traits in insider threat frameworks, to our knowledge no known studies have empirically investigated the relationships between the dark triad traits (individually or collectively) and insider cyber sabotage. The findings of our research may help enhance and extend existing models and frameworks of insider threat behavior. Additionally, the findings may contribute to empirically validating rulesets in technical systems and traits used in insider threat training and awareness programs.

Machiavellians engage in bad behaviors for some gain, narcissists engage in bad behaviors because they are only concerned with themselves, and psychopaths behave badly for the thrill, regardless of the risk to themselves or an organization.

Background

Insider threats. Insiders represent greater threats to organizations than outsiders due to their access to organizational information and information systems, especially when coupled with their advanced organizational knowledge and the trust that is often afforded to them. Insider threats exist when trusted current or former organizational members act in ways that expose the organization to risk.[9] Inappropriate insider behavior not only threatens organizational resources, it may put the survival of the organization at risk.

When discussing insider threats, it is useful to distinguish between malicious and unintentional threats. Not surprisingly, the key difference is intent. Unintentional threats come from actions (or inactions) undertaken without any malicious intent. Using an easy-to-guess password or responding to a phishing email are examples of unintentional threats. In contrast, malicious threats come from intentional acts. The CERT National Insider Threat Center (NITC) defines a malicious insider threat as "a current or former employee, contractor, or business partner who has or had authorized access to an organization's network, system, or data and intentionally exceeded or misused that access in a manner that negatively affected the confidentiality, integrity, or availability of the organization's information or information systems."[22] The research presented here pertains to malicious, rather than unintentional, insider threats.

Malicious insider threats are often described by the nature of the crime or abuse.[7] For example, a common categorization of malicious insider threats includes espionage, cyber sabotage, fraud, and theft of intellectual property.[1,20] Cyber sabotage, or the infliction of harm on some area of an organization using technology,[4,20] can result in particularly significant and diversified damage to that organization.[5]

There are numerous methods for dealing with the threat of insider sabotage. These include technical and administrative preventive and deterrent mitigation techniques. Technical approaches include user and enterprise level systems for detection focusing on monitoring of cyber data.[9] These systems include capabilities for collection, storage, analysis, and reporting based on activities and actions of individuals. Administrative controls include training and awareness programs, security policies, and processes that include securing system access paths upon precipitating events, such as demotion or termination.[20] Our research can be used to enhance both technical and administrative approaches, as discussed later.

Personality factors, particularly pathological personality traits, have been cited as one of three essential factors predisposing individuals to malicious insider threat behavior (along with opportunity and states of crisis).[23] Past insider threat cases have noted key personal predispositions as precursors to espionage and cyber sabotage. These predispositions include an unusual need for attention, sense of entitlement, arrogance, impulsivity, lack of conscience, and lack of empathy.[1] These key characteristics reported of convicted insiders are also elements of the dark triad of personality.

The underlying psychology of individual threat agents lies at the heart of the insider threat problem. Although technical controls are helpful in mitigating harmful behaviors, they are insufficient. As experience repeatedly demonstrates, insider threat behaviors occur in spite of sophisticated technical security mechanisms. One reason for this is that these mechanisms typically detect threat activity only after the activity occurs. In addition, clever bad actors are often able to circumvent technical security controls. On the other hand, common administrative controls such as security policies and associated sanctions may not account for the underling psychology of malevolent individuals. As we explain later, such individuals may ignore policies and be unmoved by the risks associated with potential sanctions. Because of this, it is important to understand the traits of atypical, malevolent insiders. Thus, neither technical nor administrative controls alone can address the malicious insider threat problem. There needs to be an advanced holistic approach that considers insiders' psychological factors. Fair and trustworthy algorithms to support the advanced systems depend on empirical evidence derived from rigorous studies.

Dark triad. There are numerous models of personality. Perhaps the most commonly known is the five-factor model, which consists of five constructs of personality that are robust across cultures.[2] However, according to some, the five-factor model fails to fully account for individual differences in personality-related behaviors, particularly when related to antisocial behaviors.[22] Recent research addresses this weakness by adding traits that represent socially malevolent behavior. The dark triad of personality represents a set of personality characteristics that are only partially accounted for by the five-factor model.[22]

The dark triad consists of three socially averse personality traits—Machiavellianism, narcissism, and psychopathy.[15] All three of these exhibit a socially malevolent character: self-promotion, emotional coldness, duplicity, and aggressiveness. However, the three traits exhibit these tendencies to different degrees. Further, all three are to some extent manipulative, Machiavellianism most markedly so. Machiavellianism is a manipulative personality characterized by interpersonal relationship strategies oriented toward manipulation, self-interest, and deception. Those with Machiavellian personalities are primarily concerned with self-interest and are typically unconcerned about others beyond how they can serve that self-interest. Not surprisingly, Machiavellianism is negatively correlated with empathy. Narcissism is characterized by a general sense of superiority, grandiosity, entitlement, and dominance. Narcissists focus on themselves and have an inflated self-view. As is the case with Machiavellianism, narcissism is negatively correlated with empathy. Psychopathy is characterized by an arrogant, deceitful approach to relationships, along with high impulsivity and thrill-seeking, and irresponsible behavior. In addition, psychopathic personalities are deficient in affect, and exhibit low empathy and low anxiety. Table 1 summarizes key characteristics the dark triad personalities.

While all three dark triad personalities are socially malevolent, they differ in how social interactions and others are viewed. All three are unconcerned with any potential negative impacts their behaviors may have on others. Machiavellians seem to be concerned about how they can manipulate and use others to achieve personal goals, without consideration of how others might be negatively affected. Machiavellians will manipulate and exploit others, but with some goal in mind. If the manipulation and exploitation is unlikely to advance the Machiavellian's goal, the Machiavellian is unlikely to bother.

Narcissists are highly self-focused. They will engage in malevolent behavior, but that behavior may be due to the narcissist's sense of grandiosity, entitlement, and superiority rather than an explicit desire to do harm or negatively impact others. Narcissists do not care about impacts on others because they view others as unimportant. Narcissists are less likely than Machiavellians or psychopaths to consciously engage in behavior that harms others. Narcissists engage in such behaviors because they do not think of others.

A more complex set of factors lead to malevolent behaviors among those with psychopathic personalities. These

Table 1. Dark Triad traits.[1,11,15,22]

Key Characteristics	Machiavellianism	Narcissism	Psychopathy
Duplicity	×	×	×
Self-promotion	×	×	×
Aggressiveness	×	×	×
Interpersonal coldness	×	×	×
Tendency to manipulate and exploit others	×	×	×
Sense of superiority	×	×	×
Low empathy	×	×	×
Callousness/lack of conscience	×	×	×
Attention to reputation	×	×	
Cynical world view	×		
Strategic calculation	×		
Sense of entitlement		×	
Sense of grandiosity		×	
Ego-reinforcement all-consuming motive		×	
Thrill-seeking			×
Low anxiety			×
Lack of impulse control			×

individuals are not focused on the end goal that drives the Machiavellian. They engage in malevolent behavior for the thrill. Further, due to their low levels of anxiety they are not concerned about "getting caught." These factors, when combined with deficient affect and low empathy, make for a dangerous combination. To summarize, Machiavellians engage in bad behaviors for some gain, narcissists engage in bad behaviors because they are only concerned with themselves, and psychopaths behave badly for the thrill, regardless of the risk to themselves or an organization.

Dark triad personality traits can predict and explain workplace behavior.[10] This thinking has been applied after the fact to explain insider threat incidents. Numerous insiders involved in well-known threat cases displayed personality traits similar to those included in the dark triad. While not labeled as such, current insider threat research identifies several dark triad traits as being related to insider threats, including a sense of entitlement and lack of empathy.[19] However, to date no known study has used the dark triad as a framework for examining insider cyber sabotage.

Despite the emergence of dark triad personality traits in insider threat cases, there is a lack of systematic empirical research into the relationship between personality traits and insider threat behavior. Our study addresses this issue by empirically demonstrating the link between dark triad traits and intentions to engage in an insider threat behavior (cyber sabotage).

Methods and Results

The study was conducted using the Amazon Mechanical Turk (MTurk) marketplace to deploy an experimental vignette to a sample of working professionals from a variety of technology, healthcare, manufacturing, finance, academic, and service-oriented organizations. The vignette described an event in which an employee, who had recently been denied a raise, accessed a restricted database that had been left open inadvertently and discovered that several peers had 20% higher salaries than the employee. The employee copied the database and posted it to the company's website. Subjects were asked to put themselves in the place of the employee when answering questions regarding the sabotage performed by the employee. A total of 768 usable observations were obtained after removing cases with missing data. The participants ranged in age from 18 to 73 years, with mean age of 34.7 years. The sample was 42% female and 58% male. All participants were employed, with a mean tenure in their current position of 6.08 years.

We used previously validated scales to measure the dark triad traits[11] and a previously validated scale for revenge, which was adapted to the vignette, to measure intentions.[3] All scale items were measured on a seven-point scale. The relationships were tested using Mplus Version 7 software to run covariance based latent variable modeling using weighted least squares means and adjusted variances (WLSMV) estimation for categorical data using a polychoric correlation matrix. All fit indices met the minimum requirements for interpretations of results. Validity and reliability were satisfactory.

The results of model testing showed that the relationships between each of the dark triad traits and intentions to engage in insider threat behavior were positive and significant, as shown in Table 2. The strength of the relationships varied across the traits. Psychopathy had the strongest relationship ($\beta = 0.559$, $p < 0.001$), followed by Machiavellianism ($\beta = 0.379$, $p < 0.001$) and narcissism ($\beta = 0.286$, $p < 0.001$).

To further examine the relationship between the dark triad and intent, we created an index of the dark triad by computing a mean of an individual's score on all three traits. For example, an individual who scored five on Machiavellianism, a four on psychopathy, and a three on narcissism had a dark triad score of four. This measure gives a more holistic view of the relationship between the dark triad concept and malevolent intent. After we computed the dark triad index score, we prepared a scatter plot with intent on the *x*-axis and the dark triad index score on the *y*-axis. Then we used color to represent the psychopathy score, and symbol size to indicate the Machiavellianism score. Narcissism is shown by the three symbols; with high, moderate, and low narcissism scores shown by the circle, plus sign, and square respectively. We defined the three different levels according to the scores distance from the mean, with plus or minus one standard deviation from the mean being high and low respectively.

Table 2. Results.

Dark Triad Trait	Beta	P-value
Machiavellianism	0.379	< 0.001
Narcissism	0.286	< 0.001
Psychopathy	0.559	< 0.001

The right-hand side of the plot may be thought of as the *danger zone*, as these individuals are above the intention midpoint. There is an interesting contrast between the lower- and upper-right quadrants. The lower-right quadrant is the least populated quadrant. One interpretation of this is that, generally speaking, those who have low dark triad scores are unlikely to have malevolent intentions. Further, no individual with a dark triad score of less than approximately 1.75 is on the higher end of the intent scale. In contrast, the upper-right quadrant has numerous circles. Circles in this quadrant are relatively large, indicating high Machiavellianism scores, and tend to be red, rather than blue, indicating relatively high psychopathy scores. Taken together, these results seem to indicate meaningful relationships between dark triad scores and intention to commit insider cyber sabotage.

We must caution that this visual analysis is exploratory. To our knowledge, there is no established method for creating a formal dark triad score. Further, there is no theoretically or empirically established rationale for causal relationships among the dark triad traits. For these reasons, we are reluctant to perform statistical tests on the relationship between dark triad scores and intentions; such tests may imply a higher level of precision that we can claim.

Discussion

Our results clearly show the link between the dark triad and intentions to engage in malicious insider threat behavior. While there are strong, positive relationships between each of the three dark triad traits and intentions, the strength of the relationships varied across the traits. Psychopathy had the strongest relationship, followed by Machiavellianism and narcissism. In this section, we discuss these results, including implications, recommendations for practice, and avenues for future research.

Before discussing our results, it is important to understand that the presence of dark triad traits is only one potential precursor to insider threat behavior. A sound cybersecurity risk assessment system should include assessing the dark triad traits along with other individual and organizational factors.[9] The presence of dark triad traits, when found with these other factors, increases risk.

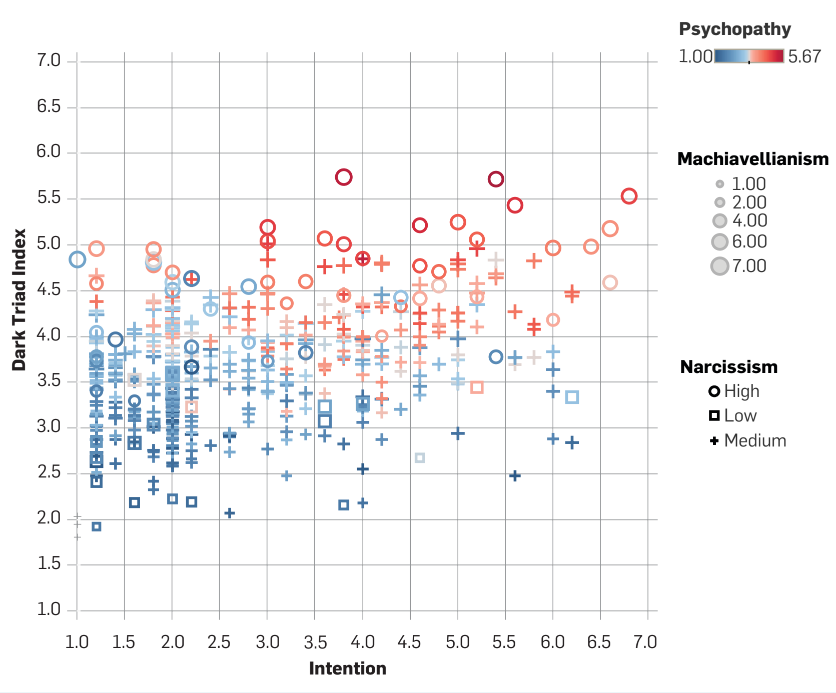
Our research answers the call for empirical validation of precursors of insider threats[9] by providing evidence of the relationship between specific personality traits and cyber sabotage. Numerous frameworks, guidelines and awareness campaigns include personal characteristics, which demonstrates the importance of personality in cyber security risk management. However, there has been a lack of empirical evidence that can be used to guide the understanding of specific concerning personality types. The empirical evidence provided in this study should be useful for refining existing cybersecurity resources related to insider threats and refine indicator development in technical monitoring systems.

There is considerable good in organizations; most individuals are hardworking and ethical. However, our research shows that individuals who tend toward the dark triad traits are more likely to intend to engage in harmful behaviors. Hence, managers should be aware of the traits and the associated tendencies toward harmful behaviors. Further, managers should have some understanding of how to deal with the threats brought about by insiders who exhibit dark triad traits.

Recommendations

Overall, managers can be thankful for the good in their organizations and for their hard working, principled employees. However, some individuals will tend toward malevolent personality traits. Because of this, developing and employing risk management strategies directed at mitigating the potential harm these individuals may bring is important. Employing such strategies is important due to the grave damage these individuals can do through their access to information assets and systems. Our data show that most employees are not inclined toward malevolent personality traits or malicious behaviors. However, as Figure 1 illustrates, malevolent people do exist, and those individuals are significantly more likely to intend to use information technology systems to do harm. These individuals are dangerous—managers should be aware of the dangers they pose.

Figure 1. Dark triad index and insider cyber sabotage intention scatterplot.



Both administrative and technical techniques exist for mitigating the potential harm from insiders. Although the following recommendations focus on administrative controls, technical controls are equally important. Our research pertains to both administrative and technical controls. The results can provide guidance for developing and refining policy, training and awareness programs, and technical monitoring systems. All of these are important elements of a well-designed cybersecurity risk management system.

Educating managers on the dark triad traits may help them identify high-risk individuals, which is an important administrative mitigation technique. Strong hiring practices that include thorough screening through multiple interviews and pre-employment testing[14] may help avoid hiring high-risk individuals. The costs of bad hires significantly exceed the potential benefits they may bring an organization, even if these individuals are "superstars."[21] Hiring managers should follow the medical community's credo, *primum non nocere*, first do no harm,[21] and actively avoid hiring high-risk insiders.

Even the most stringent hiring practices may not prevent hiring malicious individuals. Because of this, a well-designed risk management plan should assume that high-risk insiders exist in the organization and should seek to identify such individuals. However, it is important that such practices not be illegally discriminatory. Balancing the line between prudence and practices that may be seen as discriminatory requires careful thought and planning, along with the involvement of human resource and legal specialists.[4] This can be accomplished by implementing issue-specific administrative-preventative security policies as part of an overall risk mitigation plan. Many organizations have formal processes in place for insider threat cases, and if staff is well trained, behavioral patterns associated with threats resulting from the personality traits can be recognized sooner, with formal risk-management processes triggered.[4]

Sound leadership practices may help mitigate the risk from malevolent insiders. For example, psychopathic or Machiavellian individuals who perceive that some people (such as managers) are allowed to break rules that others must follow may seek revenge through the misuse of information assets. Broken promises (real or perceived) may also trigger adverse behavior in those high in dark triad traits. For example, such individuals may be motivated to do harm when a promised bonus, raise, or promotion does not come to pass. The motivation may be even stronger when the promised benefit is given to someone else.

We would be remiss if we did not mention the clear ethical dilemma posed by our recommendations. Managers face a trade-off between their responsibility to protect the organization and the rights of current and potential employees. This trade-off is similar to that presented in employee monitoring, which requires balancing organizational risk and employee privacy.[12] Hence, we consider the use of personality traits as elements of screening or monitoring efforts because assessment devices have been designed for use in normal populations and there is no reason to believe the individuals are clinically impaired regarding critical functions of the job for which they were hired.[25]

Limitations and directions for future research. This study has limitations that provide future research opportunities. First, the study confirmed a relationship but was limited in the ability to establish full causality. Due to limitations with the temporal precedence criterion, only association can be concluded. Future research using laboratory experimental or longitudinal methods are called for. Second, the generalizability of our findings is limited by our use of a specific scenario that included a particular trigger and threat. Future research should consider a range of scenarios that include additional triggering events and threat responses.

One particularly important open question concerns the effectiveness of commonly used deterrents. Some common deterrents such as sanctions may not only be ineffective for psychopaths—the risk associated with the deterrent may actually be a motivator. It is likely that what management perceives as deterrents may not be viewed as such by malevolent individuals. Another important goal for future research is to gain a better understanding of what motivates individuals high on the dark triad traits to engage in harmful behaviors. Are there particular events that lead dark triad individuals to engage in malevolent behaviors? A related question concerns how the dark triad traits influence cybersecurity decision making. Finally, researchers should seek to understand the extent to which dominant behavioral security theories apply to those exhibiting the dark triad traits, particularly in the context of malicious insider cyber security threats.

Conclusion

This study establishes the relationships between dark triad personality traits and intentions to commit insider cyber sabotage. We found strong, significant relationships between the dark triad traits and such intentions. Dark triad individuals have a propensity toward malicious insider threat acts. Although it is ideal to avoid hiring individuals exhibiting high risk traits, the difficulty in identifying such individuals during the hiring process means that they likely exist in most organizations. When it comes to malicious insider threat risk mitigation, a false positive is much better than a false negative.

Acknowledgments

Financial support was provided by the Dr. Jan Clark Memorial Research Fund and the Herbert McElveen Professorship made available through the state of Louisiana Board of Regents Support Funds and Mary Nell Condren.

References

1. Band, S.R., Cappelli, D.M., Fischer, L.F., Moore, A.P., Shaw, E.D. and Trzeciak, R.F. Comparing Insider IT Sabotage and Espionage: A Model-Based Analysis. Technical Report #CMU/SEI-2006-TR-026. Carnegie Mellon University Software Engineering Institute Pittsburgh, PA.
2. Barrick, M.R. and Mount, M.K. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology 44*, 1 (1991), 1–26.
3. Bradfield, M. and Aquino, K. 1999. The effects of blame attributions and offender likableness on forgiveness and revenge in the workplace. *J. Management 25*, 5 (1999), 607–631.
4. Cappelli, D. An unaddressed threat to critical infrastructure and national security: Insider cyber sabotage. 2018; https://bit.ly/2CpdphW.
5. Clark, J.W. Threat from within: Case studies of insiders who committed information technology sabotage. In *Proceedings of the 11th Intern. Conf. Availability, Reliability and Security* (Salzburg, Austria, Aug. 2016), 414–422.
6. CNBC. Elon Musk emails employees about "extensive and damaging sabotage" by employee. 2018; https://cnb.cx/2YnYgGr.
7. Greitzer, F.L., Frincke, D.A. and Zabriskie, M. Social/ethical issues in predictive insider threat monitoring. *Information Assurance and Security Ethics in Complex Systems: Interdisciplinary Perspectives.* Information Science Reference, 2010, 132–161.
8. Greitzer, F.L., Purl, J., Becker, D.E. (Sunny), Stitcha, P.J. and Leong, Y.M. Modeling expert judgments of insider threat using ontology structure: Effects of individual indicator threat value and class membership. In *Proceedings of the 52nd Hawaii Intern. Conf. System Sciences* (Maui, HI, USA, 2019), 3202–3211.
9. Greitzer, F.L., Purl, J., Leong, Y.M. and Sticha, P.J. Positioning your organization to respond to insider threats. *IEEE Engineering Management Review 47*, 2 (Jun. 2019), 75–83.
10. Harrison, A., Summers, J. and Mennecke, B. The effects of the dark triad on unethical behavior. *J. Business Ethics 153*, 1 (Nov. 2018), 53–77.
11. Jones, D.N. and Paulhus, D.L. Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment 21*, 1 (2014), 28–41.
12. Kiser, A.I.T., Porter, T. and Vequist, D. Employee monitoring and ethics: Can they co-exist? *Intern. J. Digital Literacy and Digital Competence 1*, 4 (Oct. 2010), 30–45.
13. Liang, N., Biros, D.P. and Luse, A. An Empirical Validation of Malicious Insider Characteristics. *J. Management Information Systems 33*, 2 (Apr. 2016), 361–392.
14. Montealegre, R. and Cascio, W.F. Technology-driven changes in work and employment. *Commun. ACM 60*, 12 (Nov. 2017), 60–67.
15. Paulhus, D.L. and Williams, K.M. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *J. Research in personality 36*, 6 (2002), 556– 563.
16. Sanders, G.L., Upadhyaya, S. and Wang, X. Inside the Insider. *IEEE Engineering Management Review. 47*, 2 (Jun. 2019), 84–91.
17. Schultz, E.E. A Framework for understanding and predicting insider attacks. *Computers & Security 21*, 6 (2002), 526–531.
18. Shaw, E. and Sellers, L. Application of the critical-path method to evaluate insider risks. *Internal Security and Counterintelligence 59*, 2 (2015), 1–8.
19. Shaw, E.D., Post, J.M. and Ruby, K.G. Inside the mind of the insider. *Security Management 43*, 12 (Dec. 1999), 34–44.
20. Software Engineering Institute. The CERT Insider Threat Center. *Common Sense Guide to Mitigating Insider Threats, Fifth Edition.* Technical Report #CMU/SEI-2015-TR-010. SEI, Carnegie Mellon University.
21. Torres, N. It's better to avoid a toxic employee than hire a superstar. *Harvard Business Review*, 2016.
22. Veselka, L., Schermer, J.A. and Vernon, P.A. The dark triad and an expanded framework of personality. *Personality and Individual Differences 53*, 4 (Sep. 2012), 417– 425.
23. Wilder, D.U.M. The psychology of espionage and leaking in the digital age. *Studies in Intelligence 61*, 2 (2017), 1–36.
24. Willison, R. and Warkentin, M. 2013. Beyond deterrence: An expanded view of employee computer abuse. *MIS Q. 37*, 1 (2013), 1–20.
25. Wu, J. and Lebreton, J.M. Reconsidering the dispositional basis of counterproductive work behavior: The role of aberrant personality. *Personnel Psychology 64*, 3 (Sep. 2011), 593–626.

Michele Maasberg (michele.maasberg@jhuapl.edu) is a Cyber Security Scientist at the Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA; https://orcid.org/0000-0003-4306-0559.

Craig Van Slyke (vanslyke@latech.edu) is the Mike McCallister Eminent Scholar Chair in Information Systems at Louisiana Tech University, Ruston, LA, USA; https://orcid.org/0000-0003-3924-1859.

Selwyn Ellis (ellis@latech.edu) is Balsley-Whitmore Endowed Professor, an associate professor, department head, and Interim Associate Dean of Graduate Programs at Louisiana Tech University, Ruston, LA, USA; https://orcid.org/0000-0002-2816-8441.

Nicole Beebe (nicole.beebe@utsa.edu) is Department Chair of Information Systems and Cyber Security and Melvin Lachman Distinguished Professor in Entrepreneurship Director of the Cyber Center for Security and Analytics at the University of Texas at San Antonio, TX, USA; https://orcid.org/0000-0002-0151-1617.

Association for Computing Machinery
STEM
Author's Rights
SIGs
Science
Open
Machine Learning
Emerging Technologies
CACM
Digital Library
Articles
Research
Transactions
Copyright
TechNews
Publishing
OpenSurround Service
Gold Open Access
Articles
eLearn
IT
Magazines
Cybersecurity
Proceedings
3D Printing
Community
Big Data
Robotics
Author-Izer
CHORUS
Networks
Cloud
HCI
AI
OpenTOC
Ubiquity
Journals
IoT
Technology
Access
HPC
XRDS
Software
Conferences
Computing
Green Open Access
ACM
Computer Science
Hybrid Open Access
Communications
Scholarly
Inroads
Computers in Entertainment
Queue

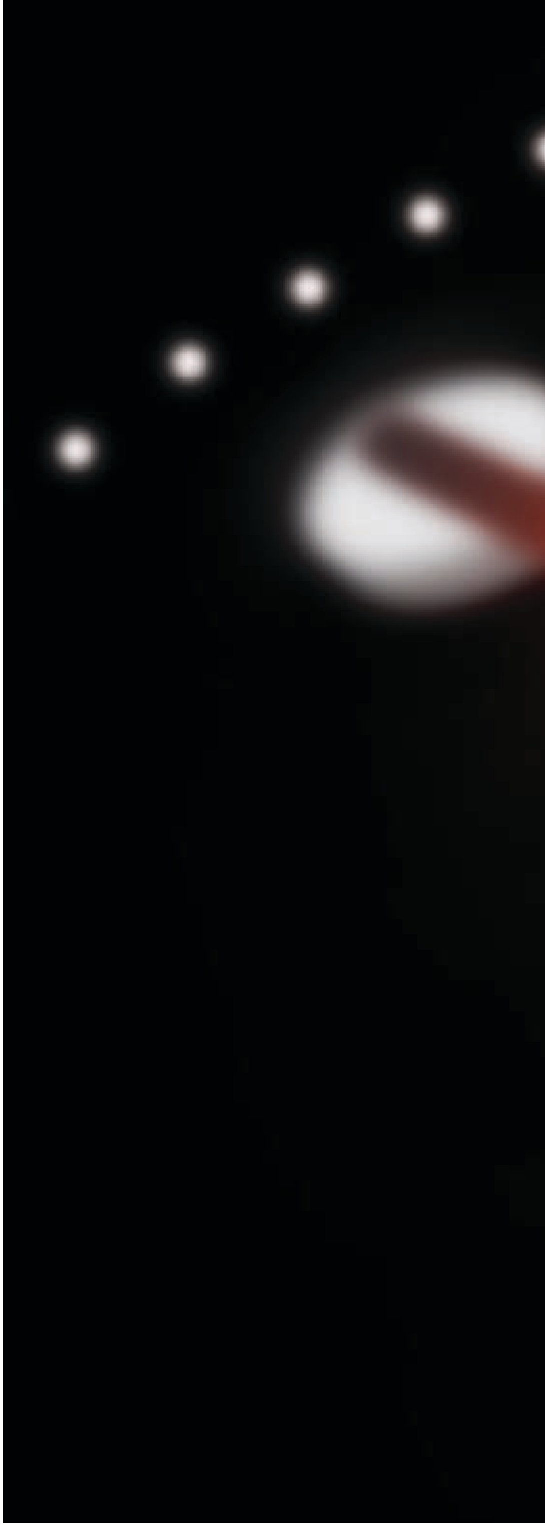
DOI:10.1145/3372135

Speed testing methods have flourished over the last decade, but none without at least some limitations.

BY NICK FEAMSTER AND JASON LIVINGOOD

Measuring Internet Speed

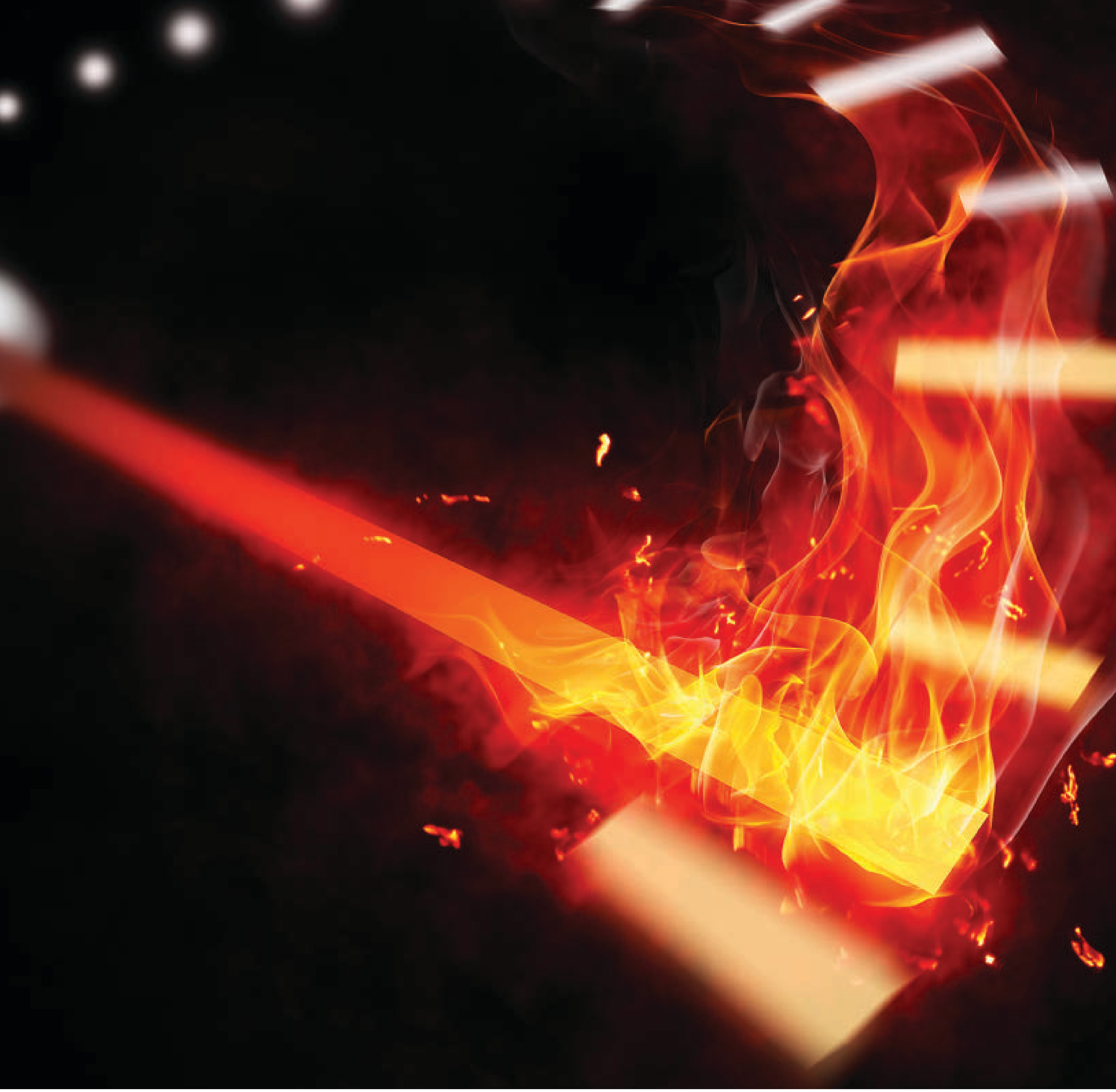
Current Challenges and Future Recommendations



VARIOUS GOVERNMENTAL ORGANIZATIONS have begun to rely on so-called Internet speed tests to measure broadband Internet speed. Examples of these programs include the Federal Communications Commission's "Measuring Broadband America" program,[7] California's CALSPEED program,[4] the U.K.'s Home Broadband Performance Program,[23] and various other initiatives in states including Minnesota,[18] New York,[19–21] and Pennsylvania.[27] These programs have various goals, ranging from assessing whether ISPs are delivering on advertised speeds to assessing potentially underserved rural areas that could benefit from broadband infrastructure investments.

The accuracy of measurement is critical to these assessments, as measurements can inform everything from investment decisions to policy actions and even litigation. Unfortunately, these efforts sometimes rely on outmoded technology, making the resulting data unreliable or misleading. This article describes the current state of speed testing tools, outlines their limitations, and explores paths forward to better inform the various technical and policy ambitions and outcomes.

Some current speed test tools were well-suited to measuring access link capacity a decade ago but are no longer useful because they made a design assumption that the Internet Service Provider (ISP) last mile access net-



work was the most constrained (bottleneck) link. This is no longer a good assumption, due to the significant increases in Internet access speeds due to new technologies. Ten years ago, a typical ISP in the United States may have delivered tens of megabits per second (Mbps). Today, it is common to have ten times faster (hundreds of megabits per second), and gigabit speeds are available to tens of millions of homes. The performance bottleneck has often shifted from the ISP access network to a user's device, home Wi-Fi network, network interconnections, speed testing infrastructure, and other areas.

A wide range of factors can influence the results of an Internet speed test, including: user-related considerations, such as the age of the device; wide-area network considerations, such as interconnect capacity; test-infrastructure considerations, such as test server capacity; and test design, such as whether the test runs while the user's access link is otherwise in use. Additionally, the typical Web browser opens multiple connections in parallel between an end user and the server to increasingly localized content delivery networks (CDNs), reflecting an evolution of applications that ultimately effects the user experience.

These developments suggest the need to evolve our understanding of the utility of existing Internet speed test tools and consider how these tools may need to be redesigned to present a more representative measure of a user's Internet experience.

IMAGE BY YUGANOV KONSTANTIN

Background

In this section, we discuss and define key network performance metrics, introduce the general principles of Internet "speed tests" and explore the basic challenges facing any speed test.

Performance metrics. When people talk about Internet "speed," they are generally talking about throughput. End-to-end Internet performance is typically measured with a collection of metrics—specifically throughput (that is, "speed"), latency, and packet loss. Figure 1 shows an example speed test from a mobile phone on a home Wi-Fi network. It shows the results of a "native" speed test from the Ookla Android speed test application[24] run in New Jersey, a canonical Internet speed test. This native application reports the user's ISP, the location of the test server destination, and the following performance metrics:

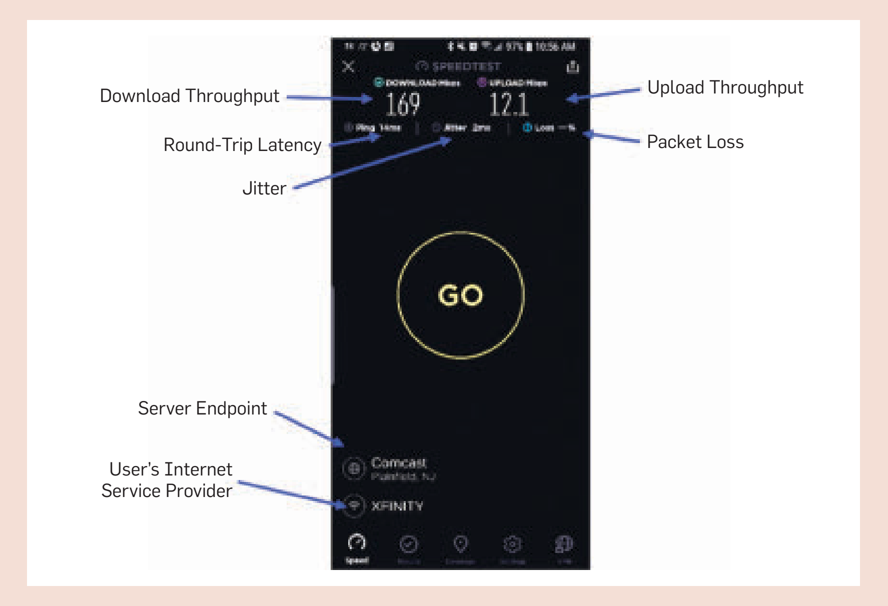
Throughput is the amount of data that can be transferred between two network endpoints over a given time interval. For example, throughput can be measured between two points in a given ISP's network, or it can be measured for an end-to-end path, such as between a client device and a server at some other place on the Internet. Typically, a speed test measures both downstream (download), from server to client, and upstream (upload), from client to server (Bauer et al.[2]) offer an in-depth discussion of throughput metrics). Throughput is not a constant; it changes from minute to minute based on many factors, including what other users are doing on the Internet. Many network-performance tests, such as the FCC test[7] and Ookla's speed test, include additional metrics that reflect the user's quality of experience.

Latency is the time it takes for a single data packet to travel to a destination. Typically, latency is measured in terms of *roundtrip latency*, since measuring one-way latency would require tight time synchronization and the ability to instrument both sides of the Internet path. Latency generally increases with distance, due to factors such as the speed of light for optical network segments; other factors can influence latency, including the amount of queueing or buffering along an end-to-end path, as well as the actual network path that traffic takes from one endpoint to another. TCP throughput is inversely proportional to end-to-end latency;[31] all things being equal, then, a client will see a higher throughput to a nearby server than it will to a distant one.

Jitter is the variation between two latency measurements. Large jitter measurements are problematic.

Packet loss rate is typically computed as the number of lost packets divided by the number of packets transmitted. Although high packet loss rates generally correspond to worse performance, some amount of packet loss is normal because a TCP sender typically uses packet loss as the feedback signal to determine the best transmission rate. Many applications such as video streaming are designed to adapt well to packet loss without noticeably affecting the end user experience, so there is no single level of packet loss that automatically translates to poor application performance. Additionally, certain network design choices, such as increasing buffer sizes, can reduce packet loss, but at the expense of latency, leading to a condition known as "buffer bloat."[3,12]

Figure 1. Example metrics from an Ookla Speedtest, a canonical Internet speed test.



Speed test principles and best practices. *Active measurement.* Today's speed tests are generally referred to as active measurement tests, meaning that they attempt to measure network performance by introducing new traffic into the network (so-called "probe traffic"). This is in contrast to *passive* tests, which observe traffic passing over a network interface to infer performance metrics. For speed testing, active measurement is the recognized best practice, but passive measurement can be used to gauge other performance factors, such as latency, packet loss, video quality, and so on.

Measuring the bottleneck link. A typical speed test sends traffic that traverses many network links, including the Wi-Fi link inside the user's home network, the link from the ISP device in the home to the ISP network, and the many network level hops between the ISP and the speed test server, which is often hosted on a network other than the access ISP. The throughput measurement that results from such a test in fact reflects the capacity of the *most constrained* link, sometimes referred to as the "bottleneck" link—the link along the end-to-end path that is the limiting factor in end-to-end throughput. If a user has a 1Gbps connection to the Internet but their home Wi-Fi network is limited to 200Mbps, then any speed test from a device on the Wi-Fi network to the Internet will not exceed 200Mbps. Bottlenecks can exist in an ISP access network, in a transit network between a client and server, in the server or server data-center network, or other places. In many cases, the bottleneck is located somewhere along the end-to-end path that is not under the ISP's or user's direct control.

Use of transmission control protocol. Speed tests typically use the Transmis-

Figure 2. TCP Dynamics.

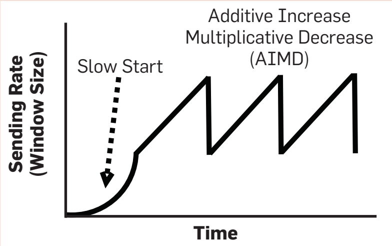
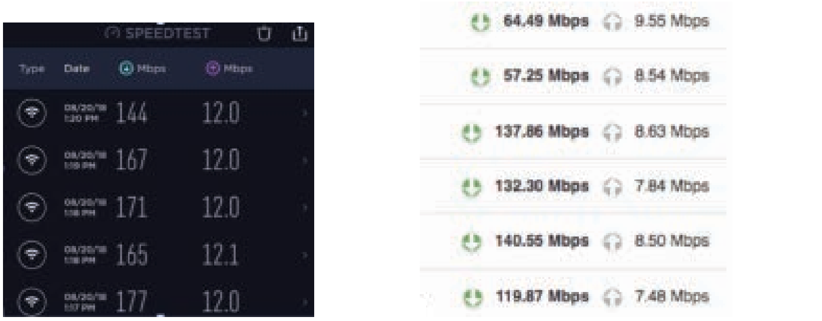


Figure 3. Successive runs of different throughput tests.



Type	Date	Mbps	Mbps
	08/20/18 1:20 PM	144	12.0
	08/20/18 1:19 PM	167	12.0
	08/20/18 1:18 PM	171	12.0
	08/20/18 1:18 PM	165	12.1
	08/20/18 1:17 PM	177	12.0

64.49 Mbps	9.55 Mbps
57.25 Mbps	8.54 Mbps
137.86 Mbps	8.63 Mbps
132.30 Mbps	7.84 Mbps
140.55 Mbps	8.50 Mbps
119.87 Mbps	7.48 Mbps

(a)
Five successive runs of Ookla Speedtest yield variable results on downstream throughput.

(b)
Internet Health Test runs in succession to six different servers. The test measures consistently lower throughput and also shows variability, both to different servers and across successive test runs.

sion Control Protocol (TCP) to measure throughput. In keeping with the nature of most Internet application transfers today—including, most notably, Web browsers—most speed tests use multiple parallel TCP connections. Understanding TCP's operation is critical to the design of an accurate speed test. Any TCP-based speed test should be: long enough to measure steady-state transfer; recognize that TCP transmission rates naturally vary over time; and, use multiple TCP connections. Figure 2 shows TCP's dynamics, including the initial slow start phase. During TCP slow start, the transmission rate is far lower than the network capacity. Including this period as part of a throughput calculation will result in a throughput measurement that is less than the actual available network capacity. If test duration is too short, the test will tend to underestimate throughput. As a result, accurate speed test tools must account for TCP slow start. Additionally, instantaneous TCP throughput continually varies because the sender tries to increase its transfer rate in an attempt to find and use any spare capacity (a process known as "additive increase multiplicative decrease" or AIMD).

Inherent variability. A speed test measurement can produce highly variable results. Figure 3 shows an illustrative example of typical variability that a speed test might yield, both for Internet Health Test (IHT) and Ookla Speedtest. These measurements were performed successively on the same Comcast connection provisioned for 200Mbps downstream and 10Mbps upstream throughput. The tests were performed in succession. Notably, successive tests yield different measurements. IHT, a Web front-end to a tool called the Network Diagnostic Test (NDT), also consistently and significantly under-reports throughput, especially at higher speeds.

Limitations of Existing Speed Tests

Existing speed tests have a number of limitations that have become more acute in recent years, largely as a result of faster ISP access links and the proliferation of home wireless networks. The most profound change is that *as network access links have become faster, the network bottleneck has moved from the ISP access link to elsewhere on the network.* A decade ago, the network bottleneck was commonly the access ISP link; with faster ISP access links, the network bottleneck may have moved any number of places, from the home wireless network to the user's device itself. Other design factors may also play a role, including how measurement samples are taken and the provisioning of the test infrastructure itself.

User-related consideration. *The home wireless network.* Speed tests that are run over a home wireless connection often reflect a measurement of the user's home wireless connection, *not* that of the access ISP, because the Wi-Fi network itself is usually the lowest capacity link between the user and test server.[1,5,16,26,28,30] Many factors affect the performance of the user's home wireless network, including: distance to the Wi-Fi Access Point (AP) and Wi-Fi signal strength, technical limitation of a wireless device and/or AP, other users and devices operating on the same network, interference from nearby APs using the same spectrum, and interference from non-Wi-Fi household devices that operate on the same spectrum (for example, microwave ovens, baby monitors, security cameras).

Many past experiments demonstrate that the user's Wi-Fi—not the ISP—is often the network performance bottleneck. Sundaresan et al. found that whenever downstream throughput exceeded 25Mbps, the user's home wireless network was almost always the bottleneck.[30] Although the study is from 2013, and both access link speeds and wireless network speeds have since increased, the general trend of home wireless bottlenecks is still prevalent.

Client hardware and software. Client types range from dedicated hardware, to software embedded in a device on the user's network, to native software made for a particular user operating system, and Web browsers. Client type has an important influence on the test results, because some may be inherently limited or confounded by user factors. Dedicated hardware examples include the SamKnows whitebox and RIPE Atlas probe. Embedded software refers to examples where the software is integrated into an existing network device such as cable modem, home gateway device, or Wi-Fi access point. A native application is software made specifically to run on a given operating system such as Android, iOS, Windows, and Mac OS. Finally, Web-based tests simply run from a Web browser. In general, dedicated hardware and embedded software approaches tend to be able to minimize the effect of user-related factors and are more accurate as a result.

Many users continue to use older wireless devices in their homes (for example, old iPads and home routers) that do not support higher speeds. Factors such as memory, CPU, operating system, and network interface card (NIC) can significantly affect throughput measurements. For example, if a user has a 100Mbps Ethernet card in their PC connected to a 1Gbps Internet connection, their speed tests will never exceed 100Mbps and that test result cannot be said to represent a capacity issue in the ISP network; it is a device limitation. As a result, many ISPs document recommended hardware and software standards,[33] especially for 1Gbps connections. The limitations of client hardware can be more subtle. Figure 4 shows an example using iPhone released in 2012–2015. This shows that any user with an iPhone 5s or older is unlikely to reach 100Mbps, likely due to the lack of a newer 802.11ac wireless interface.

Figure 4. Distribution of download speeds across different device types. Older devices do not support 802.11ac, so fail to consistently hit 100Mbps.

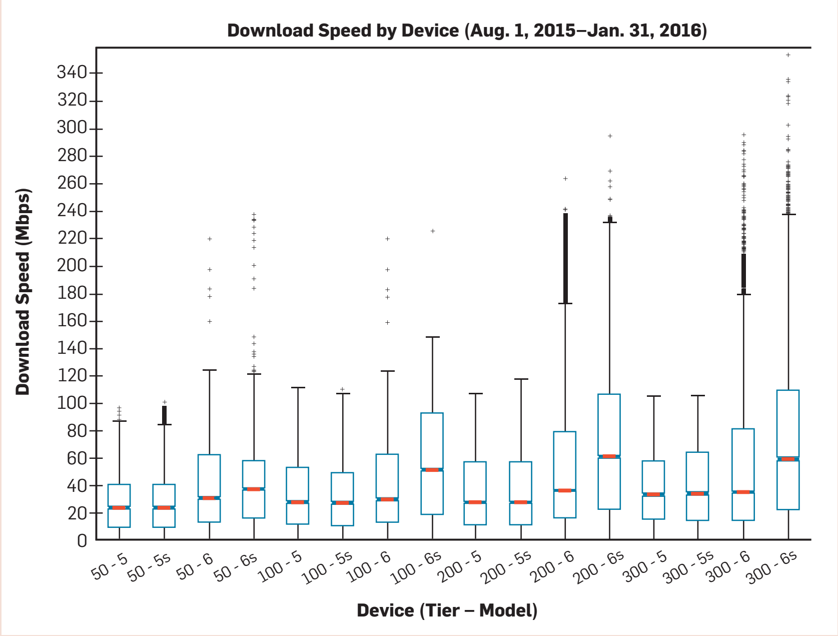
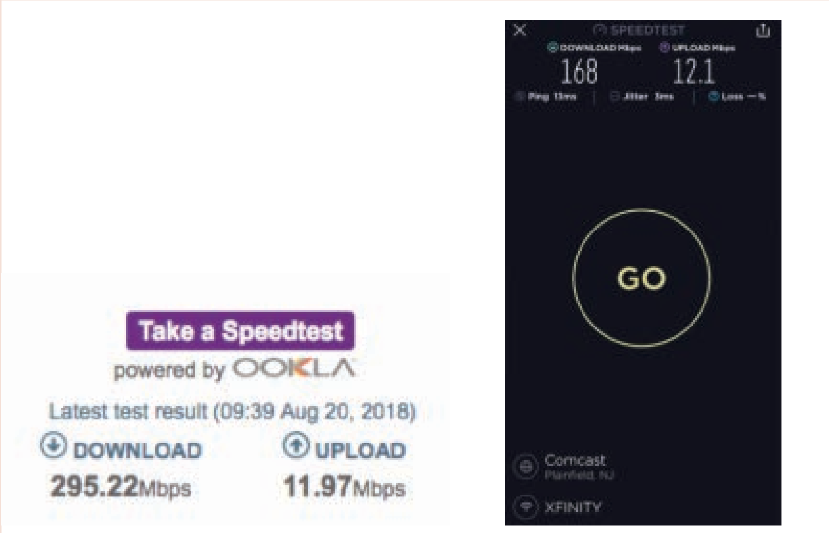


Figure 5. Two forms of Ookla Speedtest: A router-based test and native desktop test from the same home network.


Take a Speedtest

powered by OOKLA

Latest test result (09:39 Aug 20, 2018)

DOWNLOAD 295.22Mbps

UPLOAD 11.97Mbps

SPEEDTEST

DOWNLOAD Mbps 168

UPLOAD Mbps 12.1

Ping 13ms | Jitter 3ms | Loss — %

Comcast

Plainfield, NJ

XFINITY

(a)
Ookla router-based test.

(b)
Ookla native desktop test.

Router-based testing vs. device-based testing. Figure 5 shows an example of two successive speed tests. Figure 5a uses software embedded in the user's router, so that no other effects of the local network could interfere. Figure 5b shows the same speed test (such as, Ookla Speedtest), on the same network, performed immediately following the router-based test using native software on a mobile device over Wi-Fi. The throughput reported from the user's mobile device on the home network is almost half of the throughput that is reported when the speed test is taken directly from the router.

Competing "cross traffic." At any given time, a single network link is simultaneously carrying traffic from many senders and receivers. Thus, any single network transfer must share the available capacity with the competing traffic from other senders—so-called *cross traffic*. Although sharing capacity is natural for normal application traffic, a speed test that shares the available capacity with competing cross traffic will naturally underestimate the total available network capacity. Client-based speed tests cannot account for cross traffic; because the client cannot see the volume of other traffic on the same network, whereas a test that runs on the user's home router can account for cross traffic when conducting throughput measurements.

Wide-area network considerations. *Impaired ISP access network links.* An ISP's "last mile" access network links can become impaired. For example, the quality of a DOCSIS connection to a home can become impaired by factors such as a squirrel chewing through a line or a bad ground wire. Similarly, fixed wireless connections can be impaired by weather or leaves blocking the antenna. To mitigate the potential for an individual impairment unduly influencing ISP-wide results, tests should be conducted with a large number of users.

Access ISP capacity. Capacity constraints within an ISP's network can exist, whether in the access network, regional network (metropolitan area), or backbone network. Regional and backbone networks usually have excess capacity so the only periods when they may be constrained would be the result of a disaster (for example, hurricane damage) or temporary conditions such fiber cuts or BGP hijacking. Usually ISP capacity constraints arise in the last-mile access networks, which are by nature shared in the first mile or first network element, (for example, passive optical networking (PON), DOCSIS, DSL, 4G/5G, Wi-Fi, point-to-point wireless).

Transit and interconnect capacity. Another significant consideration is the connection to "transit" and "middle mile" networks. The interconnects between independently operated networks may also introduce throughput bottlenecks. As user speeds reach 1Gbps, ensuring that there are no capacity constraints on the path between the user and test server— especially across transit networks—is a major consideration. In one incident in 2013, a bottleneck in the Cogent transit network reduced NDT throughput measurements by as much as 90%. Test results improved when Cogent began prioritizing NDT test traffic over other traffic. Transit-related issues have often affected speed tests. In the case of the FCC's MBA platform, this prompted them to add servers on the Level 3 network to isolate the issues experienced with M-Lab's infrastructure and the Cogent network, and M-Labs has also added additional transit networks to reduce their reliance on one network.

Middleboxes. End-to-end paths often have devices along the path, called "middleboxes," which can affect performance. For example, a middlebox may perform load balancing or security functions (for example, malware detection, firewalls). As access speeds increase, the capacity of middleboxes may increasingly be a constraint, which will mean test results will reflect the capacity of those middleboxes rather than the access link or other measurement target.

Rate-limiting. Application-layer or destination-based rate limiting, often referred to as throttling, can also cause the performance that users experience to diverge from conventional speed tests. Choffnes et al. have developed Wehe, which detects application-layer rate limiting;[32] thus far, the research has focused on HTTP-based video streaming de-prioritization and rate-limiting. Such rate limiting could exist at any point on the network path, though most commonly it may be expected in an access network or on the destination server network. In the latter case, virtual servers or other hosted services may be priced by peak bitrate and therefore a hard-set limit on total peak bitrate or per-user-flow bitrate may exist. Web software such as Nginx has features for configuring rate limiting,[22] as cloud-based services may charge by total network usage or peak usage; for example, Oracle charges for total bandwidth usage,[25] and FTP services often enforce per-user and per-flow rate limits.[11]

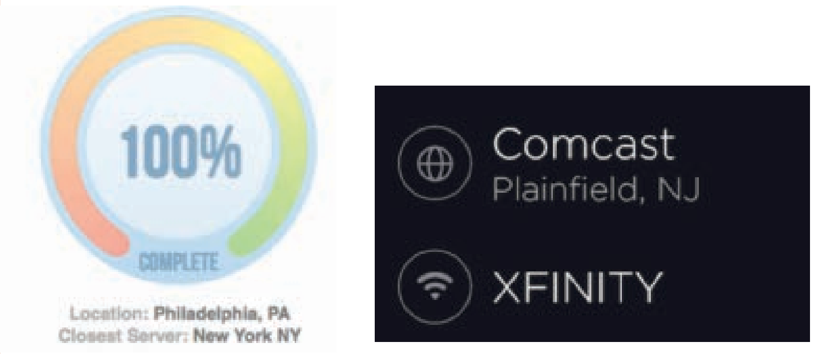
Many past experiments demonstrate that the user's Wi-Fi—not the ISP—is often the network performance bottleneck.

Rate-boosting. Rate-boosting is the opposite of rate limiting; it can enable a user to temporarily exceed their normal provisioned rate for a limited period. For example, a user may have a 100Mbps plan but may be allowed to burst to 250Mbps for limited periods if spare capacity exists. This effect was noted in the FCC's first MBA report in 2011 and led to use of a longer duration test to measure "sustained" speeds.[8] Such rate-boosting techniques appear to have fallen out of favor, perhaps partly due greater access speeds or the introduction of new technologies such as DOCSIS channel bonding.

Test infrastructure considerations. Because speed tests based on active measurements rely on performing measurements to some Internet endpoint (that is, a measurement server), another possible source of a performance bottleneck is the server infrastructure itself.

Test infrastructure provisioning. The test server infrastructure must be adequately provisioned so that it does not become the bottleneck for the speed tests. In the past, test servers have been overloaded, misconfigured, or otherwise not performing as necessary, as has been the case periodically with M-Lab servers used for both FCC MBA testing and NDT measurements. Similarly, the data center switches or other network equipment to which the servers connect may be experiencing technical problems or be subject to other performance limitations. In the case of the FCC MBA reports, at one point this resulted in discarding of data collected

Figure 6. IHT and Ookla geolocation.



(a)
Internet Health Test mistakenly locating a client in Princeton, N.J., to Philadelphia, P.A. (50+ miles away), and performing a speed test to a server in New York City.

(b)
Ookla Speedtest directing a client in Princeton, N.J., to an on-net Speedtest server in Plainfield, N.J. Ookla also allows a user to select another nearby server.

from M-Lab servers due to severe impairments.[6,9] The connection between a given datacenter and the Internet may also be constrained, congested, or otherwise technically impaired, as was the case when some M-Lab servers were single-homed to a congested Cogent network. Finally, the servers themselves may be limited in their capacity: if, for example, a server has a 1Gbps Ethernet connection (with real-world throughput below 1Gbps) then the server cannot be expected to measure several simultaneous 1Gnps or 2Gbps tests. Many other infrastructure-related factors can affect a speed test, including server storage input and output limits, available memory and CPU, and so on. Designing and operating a high scale, reliable, high-performance measurement platform is a difficult task, and as more consumers adopt 1Gbps services this may become even more challenging.[17]

Different speed test infrastructures have different means for incorporating measurement servers into their infrastructure. Ookla allows volunteers to run servers on their own and contribute these servers to the list of possible servers that users can perform tests against. Ookla uses empirical measurements over time to track the performance of individual servers. Those that perform poorly over time are removed from the set of candidate servers that a client can use. Measurement Lab, on the other hand, uses a fixed, dedicated set of servers as part of a closed system and infrastructure. For many years, these servers have been: constrained by a 1Gbps uplink; shared with other measurement experiments (recently, Measurement Lab has begun to upgrade to 10Gbps uplinks). Both of these factors can and did contribute to the platform introducing its own set of performance bottlenecks.

Server placement and selection. A speed test estimates the available capacity of the network between the client and the server. Therefore, the throughput of the test will naturally depend on the distance between these endpoints as measured by a packet's round trip time (RTT). This is extremely important, because TCP throughput is inversely proportional to the RTT between the two endpoints. For this reason, speed test clients commonly attempt to find the "closest" throughput measurement server to provide the most accurate test result and why many speed tests such as Ookla's, use thousands of servers distributed around the world. to select the closest server, some tests use a process called "IP geolocation," whereby a client location is determined from its IP address. Unfortunately, IP geolocation databases are notoriously inaccurate, and client location can often be off by thousands of miles. Additionally, latency resulting from network distance typically exceeds geographic distance, since network paths between two endpoints can be circuitous, and other factors such as network congestion on a path can affect latency. Some speed tests mitigate these effects with additional techniques. For example, Ookla's Speedtest uses IP geolocation to select an initial set of servers that are likely to be close, and then the client selects from that list the one with the lowest RTT (other factors may also play into selection, such as server network capacity). Unfortunately, Internet Health Test (which uses NDT) and others rely strictly on IP geolocation.

Figure 6 shows stark differences in server selection between two tests: Internet Health Test (which relies on IP geolocation and has a smaller selection of servers); and Ookla Speedtest (which uses a combination of IP geolocation, GPS-based location from mobile devices, and RTT-based server selection to a much larger selection of servers). Notably, the Internet Health Test not only mis-locates the client (determining that a client in Princeton, New Jersey is in Philadelphia), but it also selects a server that is in New York City, which is more than 50 miles from Princeton. In contrast, the Ookla test, which selects an on-network Comcast server in Plainfield, NJ, which is merely 21 miles away, and also gives the user the option of using closer servers through the "Change Server" option.

Test design cConsiderations. *Number of parallel connections.* A significant consideration in the design of a speed test is the number of parallel TCP connections that the test uses to transfer data between the client and server, since the goal of a speed test is to send as much data as possible and this is usually only possible with multiple TCP connections. Using multiple connections in parallel allows a TCP sender to more quickly and more reliably achieve the available link capacity. In addition to achieving a higher share of the available capacity (because the throughput test is effectively sharing the link with itself), a transfer using multiple connections is more resistant to network disruptions that may result in the sender re-entering TCP slow start after a timeout due to lost packets.

A single TCP connection cannot

typically achieve a throughput approaching full link capacity, for two reasons: a single connection takes longer to send at higher rates because TCP slow start takes longer to reach link capacity, and a single connection is more susceptible to temporarily slowing down transmission rates when it experiences packet loss (a common occurrence on an Internet path). Past research concluded that a speed test should have at least four parallel connections to accurately measure throughput.[29] For the same reason, modern Web browsers typically open as many as six parallel connections to a single server in order to maximize use of available network capacity between the client and Web server.

Test duration. The length of a test and the amount of data transferred also significantly affect test results. As described previously, a TCP sender does not immediately begin sending traffic at full capacity but instead begins in TCP slow start until the sending rate reaches a pre-configured threshold value, at which point it begins AIMD congestion avoidance. As a result, if a transfer is too short, a TCP sender will spend a significant fraction of the total transfer in TCP slow start, ensuring the transfer rate will fall far short of available capacity. As access speeds increase, most test tools have also needed to increase test duration.

Throughput calculation. The method that tests use to calculate results appears to vary widely; often this method is not disclosed. Tests may discard some high and/or low results, may use the median or the mean, may take only the highest result and discard the rest, and so on. This makes different tests difficult to compare. Finally, some tests may include all of the many phases of a TCP transfer, even though some of those phases are necessarily at rates below the capacity of a link:

- the slow start phase at the beginning of a transfer (which occurs in every TCP connection);
- the initial "additive increase" phase of the TCP transfer when the sender is actively increasing its sending rate but before it experiences the first packet loss that results in multiplicative decrease;
- any packet loss episode which results in a TCP timeout, and subsequent re-entry into slow start

Estimating the throughput of the link is not as simple as dividing the amount of data transferred by the total time elapsed over the course of the transfer. A more accurate estimate of the transfer rate would instead measure the transfer during steady-state AIMD, excluding the initial slow start period. Many standard throughput tests, including the FCC/SamKnows test, omit the initial slow start period. The Ookla test implicitly omits this period by discarding low-throughput samples from its average measurement. Tests that include this period will result in a lower value of average throughput than the link capacity can support in steady state.

Self-selection bias. Speed tests that are initiated by a user suffer from self-selection bias:[14] many users initiate such tests only when they are experiencing a technical problem or are reconfiguring their network. For example, when configuring a home wireless network, a user may run a test over Wi-Fi, then reposition their Wi-Fi AP and run the test again. These measurements may help the user optimize the placement of the wireless access point but, by design, they reflect the performance of the user's home wireless network, not that of the ISP. Tests that are user-initiated ("crowdsourced") are more likely to suffer from self-selection bias. It can be difficult to use these results to draw conclusions about an ISP, geographic region, and so forth.

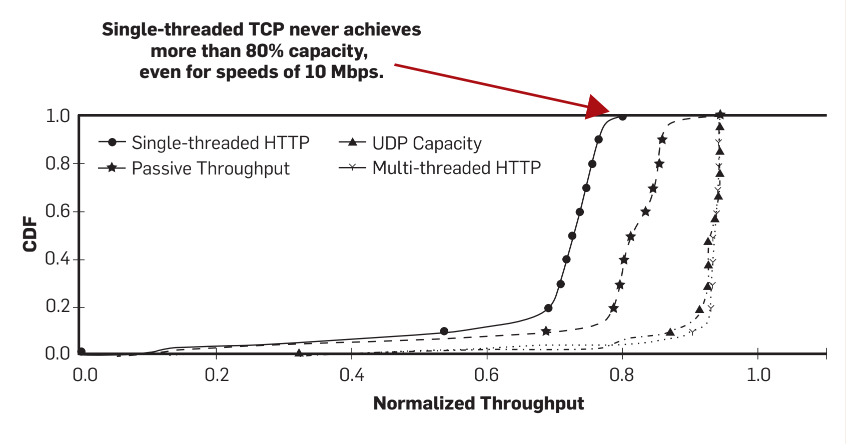
Infrequent testing. If tests are too infrequent or are only taken at certain times of day, the resulting measurements may not accurately reflect a user's Internet capacity. An analogy would be looking out a window once per day in the evening, seeing it was dark outside, and concluding that it must be dark 24 hours a day. Additionally, if the user only conducts a test when there is a transient problem, the resulting measurement may not be representative of the performance that a user typically experiences. Automatic tests run multiple times per day at randomly selected times during peak and off-peak times can account for some of these factors.

The Future of Speed Testing

Speed testing tools will need to evolve as end user connections approach and exceed 1Gbps, especially given that so many policy, regulatory, and investment decisions are based on speed measurements. As access network speeds increase and the performance bottlenecks move elsewhere on the path, speed test design must evolve to keep pace with both faster network technology and evolving user expectations. We recommend the following:

Retire outmoded tools such as NDT. NDT, also known as the Internet Health Test,[15] may appear at first glance to be suitable for speed tests. This is not the case, though it continues to be used for speed measurement despite its unsuitability and proven inaccuracy.[10] Its inadequacy for measuring access link

Figure 7. Throughput vs. number of TCP threads.



speeds has been well-documented.[2] One significant problem is that NDT still uses a single TCP connection, nearly two decades after this was shown to be inadequate for measuring link capacity. NDT is also incapable of reliably measuring access link throughput for speeds of 100Mbps or more, as we enter an era of gigabit speeds. The test also includes the initial TCP slow start period in the result, leading to a lower value of average throughput than the link capacity can support in TCP steady state. It also faces all of the user-related considerations that we discussed previously. It is time to retire the use of NDT for speed testing and look ahead to better methods.

Use native, embedded, and dedicated measurement techniques and devices. Web-based tests (many of which rely on Javascript) cannot transfer data at rates that exceed several hundred megabits per second. As network speeds increase, speed tests must be "native" applications or run on embedded devices (for example, home router, Roku, Eero, and AppleTV) or otherwise dedicated devices (for example, Odroid, Raspberry Pi, SamKnows "white box," and RIPE Atlas probes).

Control for factors along the end-to-end path when analyzing results. As we outlined earlier, many factors can affect the results of a speed test other than the capacity of the ISP link—ranging from cross-traffic in the home to server location and provisioning. As access ISP speeds increase, these limiting factors become increasingly important, as bottlenecks elsewhere along the end-to-end path become increasingly prevalent.

Measure to multiple destinations. As access network speeds begin to approach and exceed 1Gbps, it can be difficult to identify a single destination and end-to-end path that can support the capacity of the access link. Looking ahead, it may make sense to perform active speed test measurements to multiple destinations simultaneously, to mitigate the possibility that any single destination or end-to-end network path becomes the network bottleneck.

Augment active testing with application quality metrics. In many cases, a user's experience is not limited by the access network speed, but rather the performance of a particular application (for example, streaming video) under the available network conditions. As previously mentioned, even the most demanding streaming video applications require only tens of megabits per second, yet user experience can still suffer as a result of application performance glitches, such as changes in resolution or rebuffering. As access network speeds increase, it will be important to monitor not just "speed testing" but also to develop new methods that can monitor and infer quality metrics for a variety of applications.

Adopt standard, open methods to facilitate better comparisons. It is currently very difficult to directly compare the results of different speed tests, because the underlying methods and platforms are so different. Tools that select the highest result of several sequential tests, or the average of several, or the average of several tests after the highest and lowest have been discarded. As the FCC has stated:[13] "A well-documented, public methodology for tests is critical to understanding measurement results." Furthermore, tests and networks should disclose any circumstances that result in the prioritization of speed test traffic.

Beyond being well-documented and public, the community should also come to agreement on a set of standards for measuring access link performance and adopt those standards across test implementations.

References

1. Apple: Resolve Wi-Fi and Bluetooth Issues Caused by Wireless Interference, 2019. https://support.apple.com/en-us/ HT201542.
2. Bauer, S., Clark, D.D., and Lehr, W. Understanding Broadband Speed Measurements. In *Technology Policy Research Conference* (*TPRC*), 2010.
3. Bufferbloat. https://www.bufferbloat.net.
4. CALSPEED Program, 2019; http://cpuc.ca.gov/General. aspx?id=1778.
5. *Electronic Engineering Times.* Avoiding iInterference in the 2.4-GHz ISM band, 2006; https: //www.eetimes.com/document.asp?doc_id=1273359.
6. FCC. Measuring Broadband America Program (Fixed), GN Docket No. 12264, Aug. 2013; https://ecfsapi.fcc.gov/file/ 7520939594.pdf.
7. FCC: Measuring Broadband America Program; https://www.fcc. gov/general/measuring-broadband-america.
8. FCC. Measuring Broadband America. Technical report, 2011; https://transition.fcc.gov/ cgb/measuringbroadbandreport/Measuring_U.S._-_Main_Report_Full.pdf.
9. FCC. MBA Report 2014, 2014; https://www.fcc.gov/ reports-research/reports/measuring-broadbandamerica/measuring-broadband-america-2014.
10. FCC. Letter to FCC on Docket No. 17-108, 2014; https://ecfsapi.fcc.gov/file/1083088362452/fcc-17-108-reply-aug2017.pdf.
11. FTP Rate Limiting, 2019; https://forum.filezillaproject.org/viewtopic.php?t=25895.
12. Gettys, J. Bufferbloat: Dark Buffers in the Internet. In IEEE Internet Computing, 2011.
13. Google Group. MLab speed test is incorrect? 2018; https://groups.google.com/a/measurementlab.net/forum/#!topic/discuss/vOTs3rcbp38.
14. Heckman, J. J. Selection bias and self-selection. In Econometrics, pages 201–224. 1990
15. Internet Health Test, 2019. http://internethealthtest.org/.
16. Lifewire. Change the Wi-Fi channel number to avoid interference, 2018; https://www.lifewire.com/Wi-Fi-channel-numberchange-to-avoid-interference-818208.
17. Livingood, J. Measurement Challenges in the Gigabit Era, June 2018. https://blog.apnic.net/2018/06/21/measurementchallenges-in-the-gigabit-era/.
18. Minnesota.gov. CheckspeedMN; https://mn.gov/deed/programsservices/broadband/checkspeedmn.
19. New York Attorney General's Office. A.G. Schneiderman Encourages New Yorkers To Test Internet Speeds And Submit Results As Part Of Ongoing Investigation Of Broadband Providers, 2017; https://ag.ny.gov/press-release/agschneiderman-encourages-new-yorkers-testinternet-speeds-and-submit-results-part.
20. New York Attorney General's Office. Are You Getting the Internet Speeds You Are Paying For? https: //ag.ny.gov/SpeedTest.
21. New York State Broadband Program Office-Speed Test; https: //nysbroadband.ny.gov/speed-test.
22. Nginx Rate Limiting, 2019; https://www.nginx.com/blog/ rate-limiting-nginx/.
23. Ofcom. Broadband speeds: Research on fixed line home broadband speeds, mobile broadband performance, and related research; https: //www.ofcom.org.uk/research-and-data/telecomsresearch/broadband-research/broadband-speeds.
24. Ookla Speedtest, 2019; https://speedtest.net/.
25. Oracle IaaS Pricing, 2019; https://cloud.oracle.com/en_US/iaas/pricing.
26. *PC World.* Six things that block your Wi-Fi, and how to fix them, 2011; https://www.pcworld.com/article/227973/six_ things_that_block_your_Wi-Fi_and_how_to_fix_ them.html.
27. Penn State University. A Broadband Challenge: Reliable broadband internet access remains elusive across Pennsylvania, and a Penn State faculty member is studying the issue and its impact, 2018; https://news.psu.edu/story/525994/2018/06/28/research/broadband-challenge.
28. Revolution Wi-Fi. The 2.4 GHz Spectrum congestion problem and AP form factors, 2015; http://www.revolutionWi-Fi.net/ revolutionWi-Fi/2015/4/the-dual-radio-apform-factor-is-to-blame-for-24-ghz-spectrumcongestion.
29. Sundaresan, S., De Donato, W., Feamster, N., Teixeira, R., Crawford, S., and Pescape, A. broadband Internet performance: A view from the gateway. In *Proceedings of ACM SIGCOMM* (Aug. 2011), 134–145.
30. Sundaresan, S., Feamster, N., and Teixeira, R. Home network or access link? Locating last-mile downstream throughput bottlenecks. In *Proceedings of the Intern. Conf. on Passive and Active Network Measurement* (2016), 111–123.
31. Tanenbaum, A.S. et al. *Computer Networks.* Prentice Hall, 1996.
32. Wehe, 2019; https://dd.meddle.mobi.
33. Xfinity Internet Minimum System Recommendations; https://www.xfinity.com/support/articles/requirementsto-run-xfinity-internet-service.

Nick Feamster (feamster@uchicago.edu) is the Neubauer Professor in the Computer Science Department and Director of the Center for Data and Computing at the University of Chicago, IL, USA.

Jason Livingood (jason_livingood@comcast.com) is Vice President at Comcast, Philadelphia, PA, USA.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/measuring-internet-speed

research highlights

DOI:10.1145/3429950

Technical Perspective XNOR-Networks—Powerful but Tricky

To view the accompanying paper, visit doi.acm.org/10.1145/3429945 rh

By David Alexander Forsyth

YOU CAN NOW run computations on your phone that would have been unthinkable a few years ago. But as small devices get smarter, we discover new uses for them that overwhelm their resources. If you want your phone to recognize a picture of your face (image classification) or to find faces in pictures (object detection), you want it to run a convolutional neural net (CNN).

Modern computer vision applications are mostly built using CNNs. This is because vision applications tend to have a classifier at their heart—so, for example, one builds an object detector by building one classifier that tells whether locations in an image could contain an object, then another that determines what the object is. CNNs consist of a sequence of layers. Each takes a block of data which has two spatial dimensions (like an image) and one feature dimension (for an image, red, green, and blue), and then makes another such block, which it passes to the next layer. Most layers apply a convolution and a nonlinearity to their input, typically increasing the feature dimension. Some layers reduce the spatial dimension by pooling spatial windows in the input block. So, one might pool by replacing 2x2 non-overlapping windows with the largest value in that window. The final layer is usually classification by logistic regression.

CNNs yield excellent classifiers, because the training process chooses image features that are useful for the particular classification task in hand. For this to work, some layers must have quite large feature dimensions, and the network needs to have many layers (yielding a "deep" network). Deep networks produce image features that have very wide spatial support and are complicated composites. One should think of a single convolutional layer as a pattern detector; a deep network detects patterns of patterns of patterns.

All this means that CNNs tend to have a very large number of floating-point parameters, meaning that running a CNN has traditionally required a GPU (or patience!). Building networks with few parameters tends to result in classifiers that aren't accurate. But a CNN's parameters are redundant. For example, once a CNN has been trained, some procedures for compressing its parameters don't significantly affect its accuracy. CNNs can respond badly to apparently minor changes. For example, changing from single precision to double precision arithmetic can significantly affect accuracy.

How, then, to produce a CNN that is small enough to run on a mobile device, and accurate enough to be worth using? The strategies in the following paper are the best known to date. One builds a CNN where every parameter is represented by a single bit, yielding a very large reduction in size and a speedup (a binary weight network or BWN). In a BWN, layers apply binary weights to real valued inputs to get real valued outputs. Even greater improvements in size and speed can be obtained by insisting that layers accept and produce single bit data blocks (an XNOR-network). Multiplying data by weights in an XNOR-network is particularly efficient, so very significant speedups are available.

Producing a useful XNOR-network requires a variety of tricks. Pooling binary values loses more information than pooling real values, so pooling layers must be adjusted. Batch normalization layers must be moved around. Training a conventional CNN then quantizing the weights produces a relatively poor classifier. Better is to train the CNN so it "knows" the weights will be quantized, using a series of clever tricks described in the paper. It helps to adjust the labels used for training with a measure of image similarity; these refinements are adjusted dynamically throughout the training process.

These tricks result in a compression procedure that can be applied to any network architecture, with weights learned on any dataset. But compression produces a loss of accuracy. The ideal way to evaluate this procedure is to find others that produce networks of the same size and speed on the same dataset. Then the compressor that produces the smallest loss in accuracy wins. It's hard to match size and speed, but a compressor that produces the smallest loss of accuracy with acceptable size and speed is the standard to beat.

The procedures described result in accuracies much higher than is achievable with comparable methods. This work has roots in a paper that appeared in the *Proceeding of the 2016 European Conference on Computer Vision*. Since then, xnor.ai, a company built around some of the technologies in this paper, has flourished.

The technologies described mean you can run accurate modern computer vision methods on apparently quite unpromising devices (for example, a pi0). There is an SDK and a set of tutorials for this technology at https://ai2go.xnor.ai/getting-started/python. Savings in space and computation turn into savings in energy, too. An extreme example—a device that can run accurate detectors and classifiers using only solar power—was just announced (https://www.xnor.ai/blog/ai-powered-by-solar). C

David Alexander Forsyth is a professor of computer science at the University of Illinois, Urbana-Champaign, IL. USA.

DOI:10.1145/3429945

Enabling AI at the Edge with XNOR-Networks

By Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi

Abstract

In recent years we have seen a growing number of edge devices adopted by consumers, in their homes (e.g., smart cameras and doorbells), in their cars (e.g., driver assisted systems), and even on their persons (e.g., smart watches and rings). Similar growth is reported in industries including aerospace, agriculture, healthcare, transport, and manufacturing. At the same time that devices are getting smaller, Deep Neural Networks (DNN) that power most forms of artificial intelligence are getting larger, requiring more compute power, memory, and bandwidth. This creates a growing disconnect between advances in artificial intelligence and the ability to develop smart devices at the edge. In this paper, we present a novel approach to running state-of-the-art AI algorithms at the edge. We propose two efficient approximations to standard convolutional neural networks: Binary-Weight-Networks (BWN) and XNOR-Networks. In BWN, the filters are approximated with binary values resulting in 32× memory saving. In XNOR-Networks, both the filters and the input to convolutional layers are binary. XNOR-Networks approximate convolutions using primarily binary operations. This results in 58× faster convolutional operations (in terms of number of the high precision operations) and 32× memory savings. XNOR-Nets offer the possibility of running state-of-the-art networks on CPUs (rather than GPUs) in real-time. Our binary networks are simple, accurate, efficient, and work on challenging visual tasks. We evaluate our approach on the ImageNet classification task. The classification accuracy with a BWN version of AlexNet is the same as the full-precision AlexNet. Our code is available at: urlhttp://allenai.org/plato/xnornet.

1. INTRODUCTION

In recent years, the approach of using Deep Neural Networks (DNN) to create artificial intelligence has been highly successful in teaching computers to recognize[8, 11, 17, 18] and detect[4, 5, 16] objects, read text, and understand speech.[7] Such capabilities could have significant impacts on industries such a healthcare, agriculture, aerospace, transport, and manufacturing, yet to date there are limited real world applications of DNN and Convolutional Deep Neural Networks (CDNN). While there has been some progress made with virtual reality (VR by Oculus),[13] augmented reality (AR by HoloLens),[6] and smart wearable devices, the majority of applications rely on edge devices that have limited or no bandwidth, are low powered, and require the data to be stored locally for privacy and security reasons. These constraints are at odds with the current state-of-the-art CNNs and DCNNs that require large amounts of compute power and are therefore currently limited to the cloud.

CNN-based recognition systems need large amounts of memory and computational power. While they perform well on expensive, GPU-based machines, they are often unsuitable for smaller devices like cell phones and embedded electronics. For example, AlexNet,[11] one of the most well-known DNN architecture for image classification, has 61M parameters (249MB of memory) and performs 1.5B high precision operations to classify one image. These numbers are even higher for deeper CNNs for example, VGG[17] (see Section 3.1). These models quickly overtax the limited storage, battery power, and compute capabilities of smaller devices like cell phones.

In this paper, we introduce simple, efficient, and accurate approximations to CNNs by binarizing the weights and even the intermediate representations in convolutional neural networks. Our binarization method aims at finding the best approximations of the convolutions using binary operations. We demonstrate that our way of binarizing neural networks results in ImageNet classification accuracy numbers that are comparable to standard full precision networks while requiring a significantly less memory and fewer floating point operations.

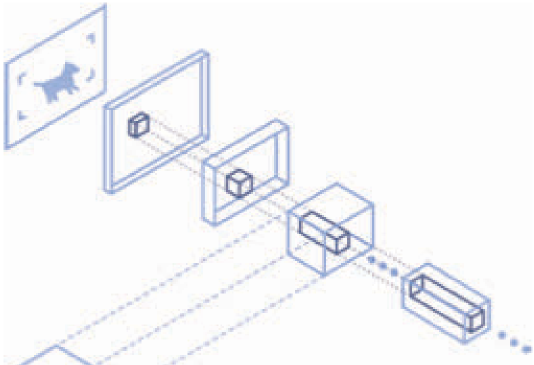
We study two approximations: Neural networks with binary weights and XNOR-Networks. In **BWN** all the weight values are approximated with binary values. A convolutional neural network with binary weights is significantly smaller ($\sim 32\times$) than an equivalent network with single-precision weight values. In addition, when weight values are binary, convolutions can be estimated by only addition and subtraction (without multiplication), resulting in $\sim 2\times$ speed up. Binary-weight approximations of large CNNs can fit into the memory of even small, portable devices while maintaining the same level of accuracy (see Sections 3.1 and 3.2).

To take this idea further, we introduce **XNOR-Networks** where both the weights and the inputs to the convolutional and fully connected layers are approximated with binary values. Binary weights and binary inputs allow an efficient way of implementing convolutional operations. If all of the operands of the convolutions are binary, then the convolutions can be estimated by XNOR and bit-counting operations.[2] XNOR-Nets result in accurate approximation of CNNs while offering $\sim 58\times$ speed up in CPUs (in terms of number of the high precision operations). This means that XNOR-Nets can enable real-time inference in devices with small memory and no GPUs (inference in XNOR-Nets can be done very efficiently on).[a]

[a] Fully connected layers can be implemented by convolution, therefore, in the rest of the paper, we refer to them also as convolutional layers.[12]

The original version of this paper is entitled "*XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks*" and was published in (European Conference on Computer Vision (ECCV) 2016.

Figure 1. We propose two efficient variations of convolutional neural networks. Binary-Weight-Networks, when the weight filters contains binary values. XNOR-Networks, when both weigh and input have binary values. These networks are very efficient in terms of memory and computation, while being very accurate in natural image classification. This offers the possibility of using accurate vision techniques in portable devices with limited resources.



		*		Operations	Memory	Computation	Accuracy Res-Net-50 (top-1)
Full precision	$\mathbb{R}$	*	$\mathbb{R}$	+ − x	1x	1x	75.7%
Binary weight	$\mathbb{R}$	*	$\mathbb{B}$	+ −	~32x	~2x	75.1%
XNOR-Networks	$\mathbb{B}$	*	$\mathbb{B}$	XNOR Bit-count	~32x	~58x	70.3%

To the best of our knowledge this paper is the first attempt to present an evaluation of binary neural networks on large-scale datasets like ImageNet. Our experimental results show that our proposed method for binarizing convolutional neural networks outperforms the state-of-the-art network binarization method of[2] by a large margin (16:3%) on top-1 image classification in the ImageNet challenge ILSVRC2012. Our contribution is two-fold: First, we introduce a new way of binarizing the weight values in convolutional neural networks and show the advantage of our solution compared to state-of-the-art solutions. Second, we introduce XNOR-Nets, a DNN model with binary weights and binary inputs and show that XNOR-Nets can obtain similar classification accuracies compared to standard networks while being significantly more efficient. Our code is available at: urlhttp://allenai.org/plato/xnornet.

2. BINARY CONVOLUTIONAL NEURAL NETWORK

To process an image for a variety of computer vision tasks, we need to pass the image through a multi-layer convolutional neural network. The major computational bottleneck is in the convolutional operations, which are combinations of simple floating point arithmetic operations. In the state-of-the-art CNN models the floating point operations are in the order of billions. This is the main reason that processing images with the state-of-the-art CNN models require GPU servers. GPUs can parallelize these huge amount of floating point operations. But GPUs are expensive and consume extensive power to run. In this paper, we are questioning floating point arithmetic operations in CNNs. We show that it is possible to reduce the precision of the parameters and the activation values for the neurons from 32 bits all the way down to a single bit. By reducing the precision we can save in memory and computation. Single bit precision enables using logical operations instead of floating point operations. Mathematically we present binary values in $\{-1, +1\}$, therefore the arithmetic operations translates to logical operations in $\{0, 1\}$. As it is shown in Figure 2 the multiplication translates to XNOR operation and addition and subtraction translate to popcount operations. These operations are natively available in the most of the commodity CPUs in edge devices and can be parallelized inside the CPU. Hence, it eliminated the need of GPU for fast computation.

2.1. Binary-Weight-Networks

In order to constrain a convolutional neural network to have binary weights, we estimate the real-value weight filter $\mathbf{W} \in \mathbb{R}^{c \times w \times h}$ using a binary filter $\mathbf{B} \in \{+1, -1\}^{c \times w \times h}$. The best approximation is easy to find; the sign values of the elements in $\mathbf{W}$. However, this approximation enforces a large amount of quantization error. To compensate this quantization error, we introduce a scaling factor $\alpha \in \mathbb{R}^{+}$ such that $\mathbf{W} \approx \alpha\mathbf{B}$. A convolutional operation can be approximated by:

$$\mathbf{I} * \mathbf{W} \approx (\mathbf{I} \oplus \mathbf{B})\, \alpha \qquad (1)$$

where, $\oplus$ indicates a convolution without any multiplication. Since the weight values are binary, we can implement the convolution with additions and subtractions. The binary weight filters reduce memory usage by a factor of $\sim 32\times$

compared to single-precision filters. In Ref.,[14] we found a closed form optimal estimation for $\mathbf{W} \approx \alpha\mathbf{B}$ by solving a constrained optimization problem. The optimal estimation of a binary weight filter can be simply achieved by taking the sign of weight values. The optimal scaling factor is the average of absolute weight values.

Training Binary-Weights-Networks. A naive solution for training the BWN could be first training a full precision model and then simply quantizing the weight values as describes above. This approach does not work. Figure 3 shows an experiment on image classification task in ImageNet dataset using a ResNet models with 50 layers. The top-1 accuracy in the second bar from the left shows the accuracy of this approach in compare with the full precision (first bar from the left). The naive quantization destroys all the information in the parameters of the network. The main challenge here is to find a set of real value weight filters that if we quantize them, we can reliably classify the categories of objects in an image. To find this set of weight filters, we adopt a modified version of gradient backpropagation algorithm.

Algorithm 1: Training a CNN with binary weights:

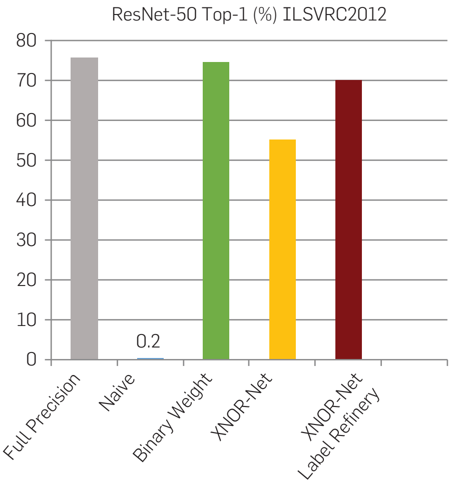
```
Input: A set of training images $\mathbf{X}$
Output: Model parameters $\mathbf{W}$
Randomly initialize $\mathbf{W}$
for $iter = 1$ to $N$ do
    Load a random image $\mathbf{X}$ from the train set
    Quantize the model parameters $\mathbf{W}$ as described above
    Forward pass the image $\mathbf{X}$ using the quantized parameters
    Compute the loss function (cross-entropy for classification)
    Backward pass to compute gradients with respect of the quantized parameters
    Update the real-value weights $\mathbf{W}$ using the gradients with a proper learning rate
```

Each iteration of training a CNN involves three steps; forward pass, backward pass and parameters update. To train a CNN with binary weights (in convolutional layers), we only quantize the weights during the forward pass and backward propagation. To compute the gradient for the sign function, we follow the same approach as.[2] For updating the parameters, we use the high precision (real-value) weights. Because, in gradient descend the parameter changes are tiny, quantization after updating the parameters ignores these changes and the training objective cannot be improved. References[2, 3] also employed this strategy to train a binary network. Algorithm 1 demonstrates a high-level schema of our procedure for training a CNN with binary weights. Once the training finished, there is no need to keep the real-value weights. Because, at inference we only perform forward propagation with the binarized weights. In Figure 3 the third bar from the left shows the accuracy of the binary weight network trained with the proposed algorithm. As it can be seen, the top-1 accuracy is as high as the full precision model while the model size is about 32× smaller.

Figure 2. Mathematically we present binary values in {−1, +1}, therefore the arithmetic operations translates to logical operations in {0, 1}. The multiplication translates to XNOR operation and addition and subtraction translate to popcount operations.

{−1,+1}	{0,1}
MUL	XNOR
ADD, SUB	Bit-Count (popcount)

Figure 3. This bar chart compares the top-1 accuracies in image classification on ImageNet challenge ILSVRC2012 using Residual Network model with 50 layers. From the left side the first bar shows the accuracy of the full precision model, the second bar shows the accuracy when the model parameters are binarized with a naive approach as discussed in Section 2.1.1, the third bar shows the accuracy when only the weights are binarized, the forth bar shows the accuracy when both weights and inputs are binarized, and the last bar shows the accuracy when the XNOR-Net model is trained with Label Refinery (see Section 2.3).



2.2. XNOR-Networks

So far, we could find binary weight filters for a CNN model. The inputs to the convolutional layers are still real-value tensors. Now, we explain how to quantize both weight filters and input tensors, so convolutions can be implemented efficiently using XNOR and bitcounting operations. This is the key element of our XNOR-Networks. In order to constrain a convolutional neural network to have binary weights and binary inputs, we need to enforce binary operands at each step of the convolutional operation. A convolution consist of repeating a shift operation and a dot product. Shift operation moves the weight filter over the input and the dot product performs element-wise multiplications between the values of the weight filter and the corresponding part of the input. If we express the dot product in terms of binary operations, convolution can be approximated using binary operations. Dot product between two binary vectors can be implemented by XNOR–bitcounting operations.[2] In Ref.,[14] we explain how to approximate the dot product between two vectors in $\mathbb{R}^n$ by a dot product between two vectors in $\{+1,$

$-1\}^n$. Similar to the binary weight approximation, we introduced scaling factor for the quantized input tensor and we found the optimal solution by solving a constrained optimization that has a closed form solution for the weight filters and the input tensors. The optimal estimation of a binary weight filter and an input tensor can be simply achieved by taking the sign of their values. The optimal scaling factors are the average of absolute values.

Next, we demonstrate how to use this approximation for estimating a convolutional operation between two tensors.

Now, using this approximation we can perform convolution between input **I** and weight filter **W** mainly using binary operations:

$$\mathbf{I} * \mathbf{W} \approx (\text{sign}(\mathbf{I}) \circledast \text{sign}(\mathbf{W})) \odot \mathbf{K}\alpha \tag{2}$$

where $\circledast$ indicates a convolutional operation using XNOR and bitcount operations and α is the scaling factor for the weight filter and **K** is a matrix of scaling factors for all of the spatial sections of the input tensor in the convolution. Note that the number of non-binary operations is very small compared to binary operations.

Training XNOR-Networks: A typical block in CNN contains several different layers. Figure 4(left) illustrates a typical block in a CNN. This block has four layers in the following order: 1-Convolutional, 2-Batch Normalization, 3-Activation, and 4-Pooling. Batch Normalization layer[9] normalizes the input batch by its mean and variance. The activation is an element-wise non-linear function (e.g., Sigmoid, ReLU). The pooling layer applies any type of pooling (e.g., max,min or average) on the input batch. For a binarized convolution the activation layer is the sign function. With the typical CNN block structure, binarization does not work. Applying pooling on binary input results in significant loss of information. For example, max-pooling on binary input returns a tensor that most of its elements are equal to +1. Moreover, in the backward pass we often observe more than one maximum that leads to uncertainty in the penalization. One may assume that switching between activation and the pooling layer will solve this issues. In this case, the input to the activation layer is real value. Max pooling will often gives us a positive tensor. Then the activation turns it into a unity matrix where most of the values are +1, which means again we loose the information for the next layer. However, this configuration does not have the penalization problem in the backward pass. Because the pooling here usually has one maximum per each window. The XNOR-Net block configuration shown in Figure 4(right), start with BatchNormalization and activation then a convolution and at the en the pooling. This configuration passes a binary input to the convolution, which generates a real-value tensor followed by a pooling which produces a tensor with mostly positive values. This tensor goes to the batch normalization in the next layer and the mean centering in the batch normalization generates negative values that when it passed to the activation, a proper binary tensor can be generated that we pass it to the next convolution.

Therefore, we put the pooling layer after the convolution. To further decrease the information loss due to binarization, we normalize the input before binarization. This ensures the data to hold zero mean, therefore, thresholding at zero leads to less quantization error. The order of layers in a block of binary CNN is shown in Figure 4(right).

Once we have the binary CNN structure, the training algorithm would be the same as Algorithm 1.

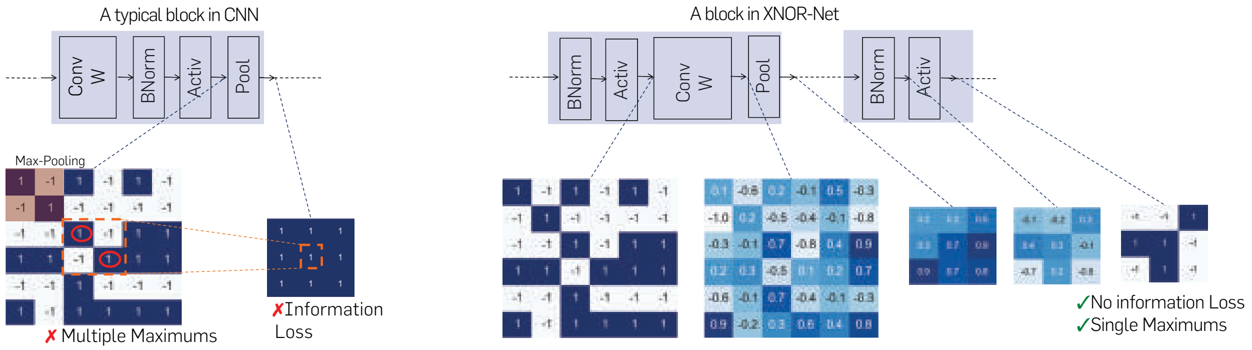
Binary Gradient: The computational bottleneck in the backward pass at each layer is computing a convolution between weight filters and the gradients with respect of the inputs. Similar to binarization in the forward pass, we can binarize the gradients in the backward pass. This leads to a very efficient training procedure using binary operations. Note that if we use the same mechanism to compute the scaling factor for quantized gradient, the direction of maximum change for SGD would be diminished. To preserve the maximum change in all dimensions, we use the maximum of the absolute values in the gradients as the scaling factor.

***k*-bit Quantization:** So far, we showed 1-bit Quantization of weights and inputs using sign(x) function. One can easily extend the quantization level to k-bits by using $q_k(x) = 2(\frac{[(2^k-1)(\frac{x+1}{2})]}{2^k-1} - \frac{1}{2})$ instead of the sign function. Where [.] indicates rounding operation and $x \in [-1, 1]$.

2.3. Improving accuracy using Label Refinery

To further improve the accuracy of the XNOR-Networks, we introduced an iterative training methods in[1] to update

Figure 4. This figure contrasts the block structure in our XNOR-Network (right) with a typical CNN (left).



ground truth labels using a visual model trained on the entire dataset. The Label Refinery produces soft, multi-category, dynamically-generated labels consistent with the visual signal. The training images are labelled with the single category. After a few iterations of label refining, the labels from which the final model is trained are informative, unambiguous, and smooth. This results in major improvements in the model accuracy during successive stages of refinement as well as improved model generalization. The last column from left in Figure 3 shows the top-1 accuracy improvement for the XNOR-Network, which is trained by Label Refinery.

3. EXPERIMENTS

We evaluate our method by analyzing its efficiency and accuracy. We measure the efficiency by computing the computational speedup (in terms of number of high precision operation) achieved by our binary convolution vs. standard convolution. To measure accuracy, we perform image classification on the large-scale ImageNet dataset. This paper is the first work that evaluates binary neural networks on the ImageNet dataset. Our binarization technique is general, we can use any CNN architecture. We evaluate AlexNet[11] and two deeper architectures in our experiments. We compare our method with two recent works on binarizing neural networks; BinaryConnect (BC)[3] and BinaryNet (BNN)[2]. The classification accuracy of our BWN version of AlexNet is as accurate as the full precision version of AlexNet. This classification accuracy outperforms competitors on binary neural networks by a large margin. We also present an ablation study, where we evaluate the key elements of our proposed method; computing scaling factors and our block structure for binary CNN. We shows that our method of computing the scaling factors is important to reach high accuracy.

3.1. Efficiency analysis

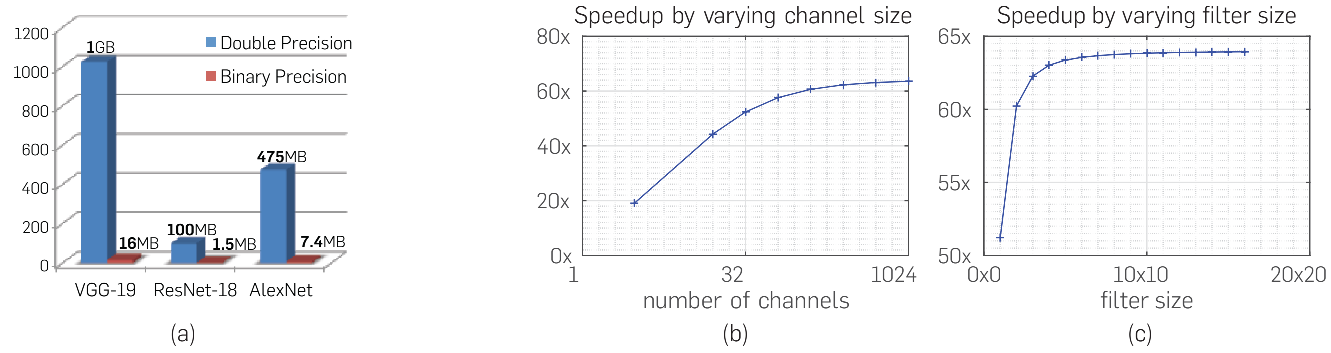
In a standard convolution, the total number of operations is $cN_{\mathrm{W}}N_{\mathrm{I}}$, where c is the number of channels, $N_{\mathrm{W}} = wh$ and $N_{\mathrm{I}} = w_{in}h_{in}$. Note that some modern CPUs can fuse the multiplication and addition as a single cycle operation. On those CPUs, BWN does not deliver speed up. Our binary approximation of convolution has $cN_{\mathrm{W}}N_{\mathrm{I}}$ binary operations and N_{I} non-binary operations. With the current generation of CPUs, we can perform 64 binary operations in one clock of CPU, therefore the speedup can be computed by $S = \frac{cN_{\mathrm{W}}N_{\mathrm{I}}}{\frac{1}{64}cN_{\mathrm{W}}N_{\mathrm{I}}+N_{\mathrm{I}}} = \frac{64cN_{\mathrm{W}}}{cN_{\mathrm{W}}+64}$.

The speedup depends on the channel size and filter size but not the input size. In Figure 5(b) and (c), we illustrate the speedup achieved by changing the number of channels and filter size. While changing one parameter, we fix other parameters as follows: $c = 256$, $n_{\mathrm{I}} = 14^2$ and $n_{\mathrm{W}} = 3^2$ (majority of convolutions in ResNet[8] architecture have this structure). Using our approximation of convolution we gain 62.27× theoretical speed up, but in our CPU implementation with all of the overheads, we achieve 58× speed up in one convolution (Excluding the process for memory allocation and memory access). With the small channel size ($c = 3$) and filter size ($N_{\mathrm{W}} = 1 \times 1$) the speedup is not considerably high. This motivates us to avoid binarization at the first and last layer of a CNN. In the first layer the channel size is three and in the last layer the filter size is 1 × 1. A similar strategy was used in.[2] Figure 5(a) shows the required memory for three different CNN architectures(AlexNet, VGG-19, ResNet-18) with binary and double precision weights. BWN are so small that can be easily fitted into portable devices. BNN[2] is in the same order of memory and computation efficiency as our method. In Figure 5, we show an analysis of computation and memory cost for a binary convolution. The same analysis is valid for BNN and BC. The key difference of our method is using a scaling-factor, which does not change the order of efficiency while providing a significant improvement in accuracy.

3.2. Image classification

We evaluate the performance of our proposed approach on the task of natural image classification. So far, in the literature, binary neural network methods have presented their evaluations on either limited domain or simplified datasets for example, CIFAR-10, MNIST, SVHN. To compare with state-of-the-art vision, we evaluate our method on ImageNet (ILSVRC2012). ImageNet has ~1.2M train images from 1K categories and 50K validation images. The images in this dataset are natural images with

Figure 5. This figure shows the efficiency of binary convolutions in terms of memory (a) and computation (b and c). (a) Figure label a is contrasting the required memory for binary and double precision weights in three different architectures(AlexNet, ResNet-18, and VGG-19). (b, c) Figure labels b and c show speedup gained by binary convolution under (b) different number of channels and (c) different filter size.



reasonably high resolution compared to the CIFAR and MNIST dataset, which have relatively small images. We report our classification performance using top-1 and top-5 accuracies. We adopt three different CNN architectures as our base architectures for binarization: AlexNet,[11] Residual Networks (known as ResNet),[8] and a variant of GoogLenet.[18]. We compare our **BWN** with **BC**[3] and our XNOR-Networks (**XNOR-Net**) with BinaryNeuralNet (**BNN**).[2] BC is a method for training a DNN with binary weights during forward and backward propagations. Similar to our approach, they keep the real-value weights during the updating parameters step. Our binarization is different from BC. The binarization in BC can be either deterministic or stochastic. We use the deterministic binarization for BC in our comparisons because the stochastic binarization is not efficient. The same evaluation settings have been used and discussed in.[2] BNN[2] is a neural network with binary weights and activations during inference and gradient computation in training. In concept, this is a similar approach to our XNOR-Network but the binarization method and the network structure in BNN is different from ours. Their training algorithm is similar to BC and they used deterministic binarization in their evaluations.

CIFAR-10: BC and BNN showed near state-of-the-art performance on CIFAR-10, MNIST, and SVHN dataset. BWN and XNOR-Net on CIFAR-10 using the same network architecture as BC and BNN achieve the error rate of 9.88% and 10.17% respectively. In this paper, we explore the possibility of obtaining near state-of-the-art results on a much larger and more challenging dataset (ImageNet).

AlexNet: Reference[11] is a CNN architecture with five convolutional layers and two fully-connected layers. This architecture was the first CNN architecture that showed to be successful on ImageNet classification task. This network has 61M parameters. We use AlexNet coupled with batch normalization layers.[9]

Train: In each iteration of training, images are resized to have 256 pixel at their smaller dimension and then a random crop of 224 × 224 is selected for training. We run the training algorithm for 16 epochs with batch size equal to 512. We use negative-log-likelihood over the soft-max of the outputs as our classification loss function. In our implementation of AlexNet we do not use the Local-Response-Normalization (LRN) layer. We use SGD with momentum = 0.9 for updating parameters in BWN and BC. For XNOR-Net and BNN we used ADAM.[10] ADAM converges faster and usually achieves better accuracy for binary inputs.[2] The learning rate starts at 0.1 and we apply a learning-rate-decay = 0.01 every four epochs.

Test: At inference time, we use the 224 × 224 center crop for forward propagation.

Figure 6 demonstrates the classification accuracy for training and inference along the training epochs for top-1 and top-5 scores. The dashed lines represent training accuracy and solid lines shows the validation accuracy. In all of the epochs our method outperforms BC and BNN by large margin (∼17%). Table 1 compares our final accuracy with BC and BNN. We found that the scaling factors for the weights (α) are much more effective than the scaling factors for the inputs (β). Removing β reduces the accuracy by a small margin (less than 1% top-1 AlexNet).

Binary Gradient: Using XNOR-Net with binary gradient the accuracy of top-1 will drop only by 1.4%.

Residual Net: We use the ResNet-18 proposed in[8] with short-cut type B.

Train: In each training iteration, images are resized randomly between 256 and 480 pixel on the smaller dimension and then a random crop of 224 × 224 is selected for training. We run the training algorithm for 58 epochs with batch size equal to 256 images. The learning rate starts at 0.1 and we use the learning-rate-decay equal to 0.01 at epochs number 30 and 40.

Test: At inference time, we use the 224 × 224 center crop for forward propagation.

Figure 7 demonstrates the classification accuracy (top-1 and top-5) along the epochs for training and inference. The dashed lines represent training and the solid lines represent inference. Table 2 shows our final accuracy by BWN and XNOR-Net.[b]

GoogLenet variant: We experiment with a variant of GoogLenet[18] that uses a similar number of parameters and connections but only straightforward convolutions, no branching. It has 21 convolutional layers with filter sizes alternating between 1 × 1 and 3 × 3.

Train: Images are resized randomly between 256 and 320 pixel on the smaller dimension and then a random

[b] Our implementation is followed by https://gist.github.com/szagoruyko/dd032c529048492630fc.

Figure 6. This figure compares the imagenet classification accuracy on top-1 and top-5 across training epochs. Our approaches BWN and XNOR-Net outperform BinaryConnect (BC) and BinaryNet (BNN) in all the epochs by large margin (∼17%).

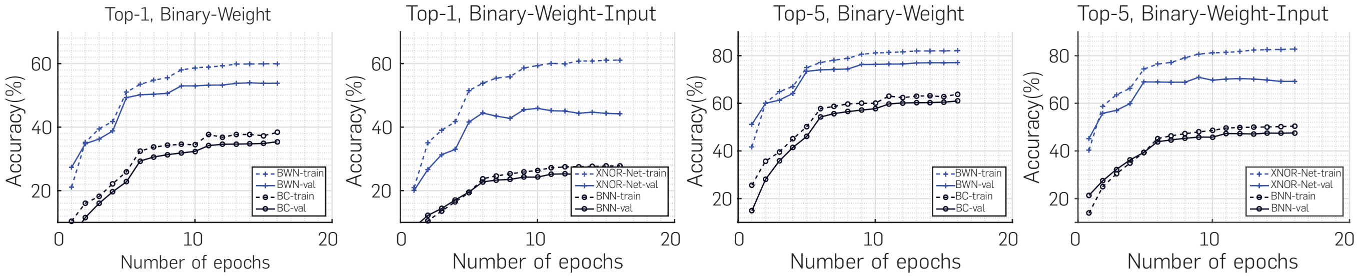


Table 1. This table compares the final accuracies (top-1–top-5) of the full precision network with our binary precision networks; Binary-Weight-Networks (BWN) and XNOR-Networks (XNOR-Net) and the competitor methods; BinaryConnect (BC) and BinaryNet (BNN).

Classification accuracy (%)									
Binary-Weight				Binary-Input-Binary-Weight				Full-precision	
BWN		BC[2]		XNOR-Net		BNN[2]		AlexNet[11]	
Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
56.8	**79.4**	35.4	61.0	**44.2**	**69.2**	27.9	50.42	56.6	80.2

Figure 7. This figure shows the classification accuracy; (a) top-1 and (b) top-5 measures across the training epochs on ImageNet dataset by Binary-Weight-Network and XNOR-Network using ResNet-18.

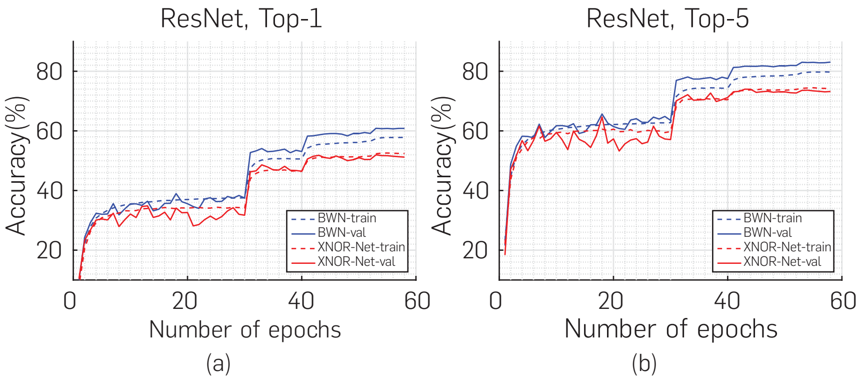


Table 2. This table compares the final classification accuracy achieved by our binary precision networks with the full precision network in ResNet-18 and GoogLenet architectures.

	ResNet-18		GoogLenet	
Network variations	**Top-1**	**Top-5**	**Top-1**	**Top-5**
Binary-Weight-Network	60.8	83.0	65.5	86.1
XNOR-Network	51.2	73.2	N/A	N/A
Full-precision-network	69.3	89.2	71.3	90.0

Table 3. In this table, we evaluate two key elements of our approach; computing the optimal scaling factors and specifying the right order for layers in a block of CNN with binary input. (a) Demonstrates the importance of the scaling factor in training Binary-Weight-Networks and (b) shows that our way of ordering the layers in a block of CNN is crucial for training XNOR-Networks. C, B, A, P stands for Convolutional, BatchNormalization, Active function (here binary activation), and Pooling respectively.

	Binary-Weight-Network	
Strategy for computing α	**Top-1**	**Top-5**
Using the scaling factor	56.8	79.4
Using a separate layer	46.2	69.5
(a)		

	XNOR-Network	
Block structure	**Top-1**	**Top-5**
C-B-A-P	30.3	57.5
B-A-C-P	44.2	69.2
(b)		

crop of 224 × 224 is selected for training. We run the training algorithm for 80 epochs with batch size of 128. The learning rate starts at 0.1 and we use polynomial rate decay, $\beta = 4$.

Test: At inference time, we use a center crop of 224 × 224.

3.3. Ablation studies

There are two key differences between our method and the previous network binarization methods; the binarization technique and the block structure in our binary CNN. For binarization, we find the optimal scaling factors at each iteration of training. For the block structure, we order the layers in a block in a way that decreases the quantization loss for training XNOR-Net. Here, we evaluate the effect of each of these elements in the performance of the binary networks. Instead of computing the scaling factor α, one can consider α as a network parameter. In other words, a layer after binary convolution multiplies the output of convolution by an scalar parameter for each filter. This is similar to computing the affine parameters in batch normalization. Table 3(a) compares the performance of a binary network with two ways of computing the scaling factors. As we mentioned in Section 2.2.1 the typical block structure in CNN is not suitable for binarization. Table 3(b) compares the standard block structure C-B-A-P (Convolution, Batch Normalization, Activation, and Pooling) with our structure B-A-C-P. (A, is binary activation).

4. CONCLUSION[c]

We introduce simple, efficient, and accurate binary approximations for neural networks. We train a neural network that learns to find binary values for weights, which reduces the size of network by $\sim$32× and provide the possibility of loading very DNN into portable devices with limited memory. We also propose an architecture, XNOR-Net, that uses mostly bitwise operations to approximate convolutions. This provides $\sim$58× speed up and enables the possibility of running the inference of state of the art DNN on CPU (rather than GPU) in real time. C

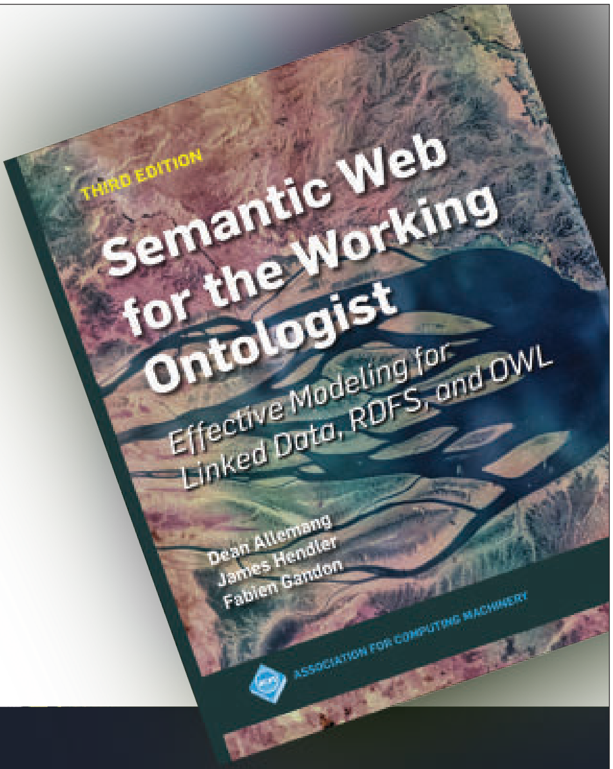
[c] We used the Darknet[15] implementation: http://pjreddie.com/darknet/imagenet/#extraction.

References

1. Bagherinezhad, H., Horton, M., Rastegari, M., Farhadi, A. Label refinery: Improving imagenet classification through label progression. arXiv preprint arXiv:1805.02641 (2018).
2. Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1. arXiv preprint arXiv:1602.02830 (2016).
3. Courbariaux, M., Bengio, Y., David, J.-P. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in Neural Information Processing Systems* (2015), 3105–3113.
4. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), 1440–1448.
5. Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), 580–587.
6. Gottmer, M. *Merging Reality and Virtuality with Microsoft Hololens*, 2015.
7. Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014).
8. He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–778.
9. Ioffe, S., Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015).
10. Kingma, D., Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
11. Krizhevsky, A., Sutskever, I., Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (2012), 097–1105.
12. Long, J., Shelhamer, E., Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 3431–3440.
13. Oculus, V. Oculus rift-virtual reality headset for 3D gaming, 2012. URL: http://www.oculusvr.com.
14. Rastegari, M., Ordonez, V., Redmon, J., Farhadi, A. XNOR-Net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision* (2016), Springer.
15. Redmon, J. Darknet: Open source neural networks in C, 2013–2016. http://pjreddie.com/darknet/.
16. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems* (2015), 91–99.
17. Simonyan, K., Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
18. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 1–9.

Mohammad Rastegari, Joseph Redmon, and Ali Farhadi ([mohammad,joe,ali]@xnor.ai), XNOR.AI, Seattle, WA, USA.

Vicente Ordonez (vicente@virginia.edu), Department of Computer Science, University of Virginia, Charlottesville, VA, USA.


THIRD EDITION
Semantic Web for the Working Ontologist
Effective Modeling for Linked Data, RDFS, and OWL
Dean Allemang
James Hendler
Fabien Gandon
ASSOCIATION FOR COMPUTING MACHINERY

DOI:10.1145/3429952

To view the accompanying paper, visit doi.acm.org/10.1145/3429948

Technical Perspective
The Future of Large-Scale Embedded Sensing

By Joseph A. Paradiso

THE DREAM OF computational material has been in the air for decades, dating at least to the Smart Matter program at Xerox PARC in the late 1990s. Inspired by the complexity of biological skin, my own team (see *Sensate Media*, *Communications*, Mar. 2005, p. 70) and others have looked to integrate distributed sensing into large flexible membranes, a trend that continues in research today (see the *IEEE 2019 Proceedings on Flexible Electronic Skin*). Most of these devices, however, are actively powered. The SATURN system described in the following paper works passively, energized essentially by static electricity generated as layers move relative to each other during vibration, hearkening perhaps to, at a smaller scale, electret microphones, which exploit charge trapped on their foil membrane to produce a vibration-dependent voltage. Using only two components—an FET and matching inductor—the authors are able to modulate the resonance of a RF antenna that can be embedded in the material and read out via passive backscatter from an external transmitter, allowing a material to work as an audio pickup without a power source.

Traditional work in this area has tended to exploit piezoelectric polymers like PVDF, which generate voltage under strain. Triboelectrics present a different approach, although provide probably an even higher source impedance that would challenge power conversion even more. SATURN sidesteps this entirely, using the generated voltage directly at the gate of the resonance-modulating FET (ironic, in that we usually work to avoid destructive static charge there—but the potentials are much lower here). Hence, the key contribution of this paper is a means of transmitting audio features from a passive triboelectric-generating material.

Remotely monitoring audio from backscatter in passive structures has a long and notorious history in electronic espionage—classic stories abound of ingenious Russian bugs built into resonant structures all over the U.S. Embassy in Moscow from over a half-century ago—microwave backscatter from cavities with a flexible surface could pick up audio across the complex (see Eric Haseltine's *The Spy in Moscow Station*, for example, Leon Theremin, famous for his free-gesture electronic musical instrument from circa 1920, is purported to be the inventor of these devices). But as this paper attests, it's back in vogue again—some recent incarnations of passive acoustic sensor backscatter can also be found in the recent work of Josh Smith's team at UW (for example, his batteryless cellphone) and my colleague Fadel Afib's self-powered underwater sonar backscatter sensor, which is acoustically interrogated instead of using radio.

The authors espouse the vision of large surface-area passive sensors that can be cheaply manufactured, perhaps by a roll-roll process, and laminated onto the walls and surfaces in our environment. There is potential competition here, however, from the opposite tack—making the sensors small and compact using MEMS and standard IC technology and embedding them into the smart surface like raisins in pudding. Looking at the application proposed here, for example, Jon Bernstein and his team at Draper Lab have recently built a passive MEMS acoustic switch that closes at a particular sonic amplitude—this could easily toggle a backscatter antenna to enable remote readout. The world also begins to see implementations of 'Smart Dust' as envisioned by Kris Pister in 2001—for example, compact stacks of bare IC die, sparsely powered by photodetectors on the top layer and talking via backscatter, such as prototyped by a University of Michigan consortium.

How we will power these sensors is an area of similar technical tension. When the power requirement is sufficiently low to warrant energy scavenging, a small, embedded battery will generally survive close its shelf life, which can approach the product life cycle of the embedded sensors and provide a more economical and practical energy solution. On the other hand, if we have a multitude of devices in our environments that must last decades, energy harvesting may be mandated, and here we will need area for photovoltaics, thermoelectrics, piezoelectrics, or even, as in this case, perhaps triboelectrics or maybe RF or inductive energy receivers. Large flat sheets provide such area, and researchers have built systems using all of these approaches (see 'The Superpowers of Super-Thin Materials,' *NYT* Jan. 7, 2020), even building sensors and electronics into fibers and fabrics, but it's not yet clear what the driving applications are. We will see flexible display 'wallpaper' in the not too distant future, but this will definitely be a powered system (this world will witness an interesting tension between photons beamed to our retinas via ubiquitous AR glasses vs light from everywhere displays). Perhaps its first market will be in building materials (for example, passively detecting dampness, strain, or temperature, after they are installed, as envisioned in my team's original 'Sensor Tape' project from 2012).

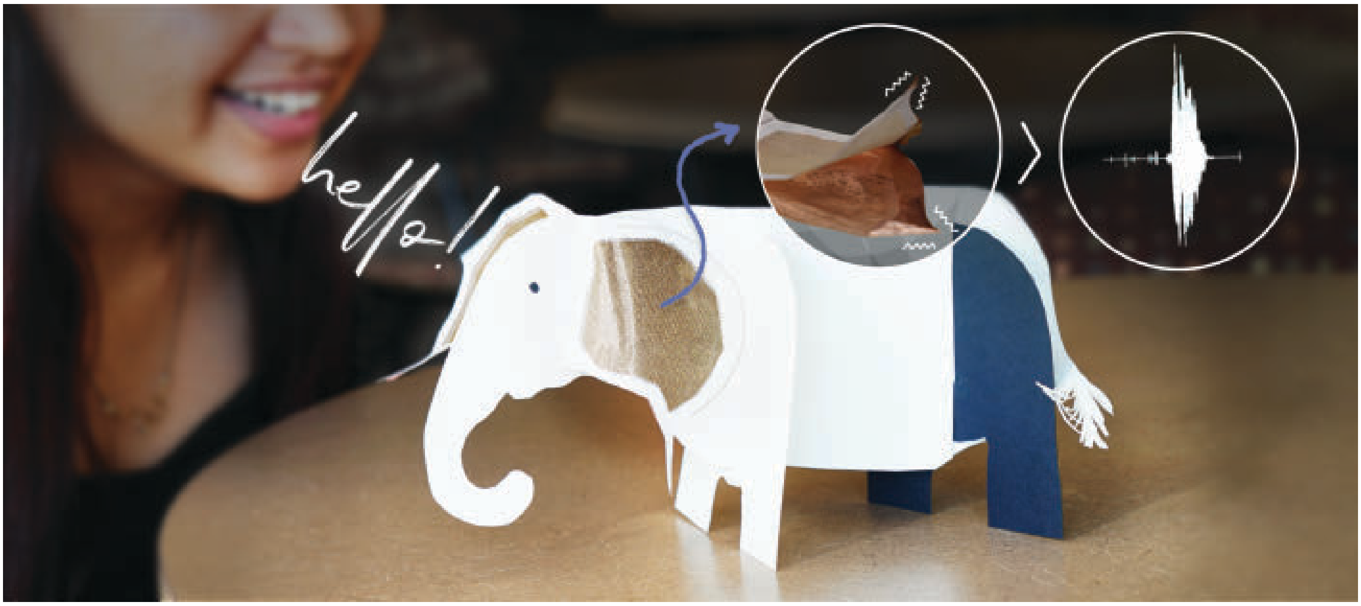
Passive sensate structures, as espoused in this paper, will enable sensing everywhere. We are already living in a world where networked sensing risks privacy behind every door—once our commonplace materials beam new streams of ubiquitous sensor data, this reaches another level, as even coarse but plentiful data can leverage potentially invasive contextual determination. The paper describes some simple ideas of physically 'opting in' with these materials, but I think the details of how privacy will be managed will be much more complex when life is enveloped with so many digital peepholes looking at us from everything. C

Joseph A. Paradiso is the Alexander W. Dreyfoos (1954) Professor and Associate Academic Head of the Program in Media Arts and Sciences at the Massachusetts Institute of Technology, Cambridge, MA, USA.

DOI:10.1145/3429948

SATURN: An Introduction to the Internet of Materials

By Nivedita Arora, Thad Starner, and Gregory D. Abowd



Sound impacts the SATURN vibration sensor, which is formed in the shape of the elephant's ear. SATURN's triboelectric components vibrate (inset), generating an electrical signal.

Abstract

We envision a new generation of computation devices, *computational materials*, which are self-sustainable, cheaply manufactured at scale and exhibit form factors that are easily incorporated into everyday environments. These materials can enable ordinary objects such as walls, carpet, furniture, jewelry, and cups to do computational things without looking like today's computational devices. Self-powered Audio Triboelectric Ultra-thin Rollable Nanogenerator (SATURN) is an early example of a computational material that can sense vibration, such as sound. SATURN can be manufactured from inexpensive components, is flexible so that it can be integrated into many different surfaces, and powers itself through the sound or vibration it is sensing. Using radio backscatter, we demonstrate that SATURN's sensed data is passively transmitted to remote computers, alleviating the need for batteries or any wired power for the material itself. The proliferation of these types of computational materials ushers an era of Internet of Materials, further blurring the distinction between the physical and digital worlds.

1. INTRODUCTION

> The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.[18]

This poetic and mostly metaphorical vision from Mark Weiser inspired nearly 30 years of ubiquitous computing research. Weiser correctly predicted an era of proliferation of many differently sized devices aimed at augmenting our human experience with technology. Today's realization of ubiquitous computing devices—smartphones, tablets, electronic whiteboards, and wearables—is still fairly easy to distinguish from everyday objects in the physical world. Sizes may vary, but there are still very distinctive characteristics of something that is computational. The Internet of Things (IoT) has tried to hide computing into more and more everyday objects, such as light bulbs, television sets, and speakers, but we are still far from a complete blurring of the physical and digital worlds. To make something computational still requires "smarts" composed of off-the-shelf integrated circuits housed in rigid modules that are packaged with existing objects. Computing is too separate from the materials of everyday objects. We propose a different direction based on Weiser's vision, that is, starting with the materials of everyday life and creating computation from there. In doing so, we propose an **Internet of Materials (IoM),** where the very materials of objects and surfaces are augmented or manufactured to have computational capabilities. This recasting

The original version of this paper is entitled "SATURN: A thin and flexible self-powered microphone designed on the principle of triboelectric nanogenerator" and was published in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.2 (60), 2018, ACM

of ubiquitous computing as "computational materials" presents three major challenges:

- **Power:** Computational materials should be self-sustainable in terms of power consumption. They should not require a wired, constant power source or battery replacements. Instead, they should be able to harvest power from the environment for operation. We would never want to plug in an everyday object, such as a cup, nor do would we ever want to worry about replacing a battery or recharging it.
- **Cost:** Computational materials should be cheap to manufacture at large scale. Despite progress in manufacturing increasingly complex and powerful integrated circuits, as dictated by Moore's law, we cannot afford to make computational materials the same way. It would be too expensive and would not scale to cover entire buildings, outdoor surfaces, or clothing.
- **Form Factor:** Computational materials should be flexible or dispersible such that they can be easily integrated into everyday objects and surfaces. The key to blurring the physical and digital worlds is to make the digital objects look and feel more like physical objects.

For several years, we have been driving our research with these three challenges. We present here one of our first successful examples of a computational material, **SATURN** (**S**elf-powered **A**udio **T**riboelectric **U**ltra-thin **R**ollable **N**anogenerator). SATURN is a demonstration of a thin and flexible multilayer material for detecting mechanical vibrations, which are ubiquitous in our everyday environment. Inspired by recent results in materials science, SATURN produces energy from mechanical vibrations using a truly ubiquitous phenomenon that occurs between any two different materials rubbing against each other. We show how this fundamental property of materials can be turned into a self-sustaining sensor and then demonstrate how other simple materials can be used to help turn the sensor into a wireless sensor.

We outline the paper in six main sections. Section 2 introduces the design and working principles of SATURN. Section 3 characterizes SATURN as a self-sustainable microphone and loud sound energy harvester. It is followed by Section 4 which demonstrates different ways of building a self-sustainable computational system leveraging SATURN. Next we discuss applications in Section 5, some of which are implemented and some are exploratory in nature. In Section 6, we discuss how this work impacts our thinking in developing computational materials.

2. SATURN: WORKING PRINCIPLES

Recent advances in materials science demonstrate the possibility of self-powered, easy-to-manufacture sensors that take advantage of the triboelectric nanogenerator (TENG) effect which converts mechanical vibrations into electrical energy.[14, 16, 17] When made in the right form factor, these mechanical energy generators could be manufactured as self-sustainable sound and vibration sensors. We use these principles for the design, fabrication, and evaluation of SATURN,[3] a flexible, self-powered sound sensor and sound energy harvester that is constructed with thin and inexpensive materials. The novelty of the work lies in creating a device that considers power, cost, and form factor as central design parameters without sacrificing signal quality (**Table 1**).

2.1. Triboelectric nanogenerator (TENG)

When two different materials come into contact and separate, or rub alongside each other, they tend to either gain or lose electrons, based on their position relative to each other in the triboelectric series.[21] This common phenomenon of exchange of electrons is called triboelectrification. The redistribution of charge creates an electric potential between the layers. If there is a conductive path between the two layers, the charge difference will balance due to electrostatic induction. Repeated contact and separation, therefore, produces an alternating current.[15] This multilayer structure, consisting of different materials that are both conductive on one side, is called a triboelectric nanogenerator (TENG).

2.2. Device design and fabrication

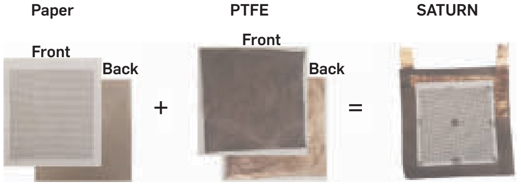
SATURN is an example of TENG and consists of two layers (Figure 1). The first is the copper that acts as a triboelectrically positive material and is coated onto paper for mechanical support. Paper is used because of its low cost, although its flexibility, light weight, and ease of perforation all support its vibration in the presence of sound. The second layer is a dielectric plastic, PTFE (polytetrafluoroethylene), which acts as a triboelectrically negative material and is coated on one side with copper using physical vapor deposition. The first and second layers are placed on top of each other, with the copper side of the paper touching the noncopper-coated side of the PTFE. The layers are anchored to each other using glue with a specific dot pattern. A potential difference is caused by vibration and is measured between the two copper-coated surfaces (see Figure 2).

SATURN's structural design is tuned to increase its electrical response across a wide frequency range. Structural

Table 1. Power generated by SATURN at 100 dB and 250 Hz sound frequency.

Size of patch	4 × 4 cm^2
Resonant frequency	255 Hz
Load impedance	0.9 MΩ
Max. V_{pp}	2.5 V
Max. power	6944 nW

Figure 1. Structural design of a SATURN microphone consisting of copper-coated paper and PTFE.



design parameters (Figure 3) such as hole size and spacing, the geometry of the patch, and the glue points attaching the two layers are varied to understand the effects on signal quality using a combination of evaluation techniques. These parameters are optimized empirically based on measured voltage generated by the microphone at a standardized decibel level and by using a mechanical model simulation that compares the separation distance between the two layers of SATURN when vibrating.

2.3. Operation

Change in air pressure due to sound vibrations causes constant contact and separation in the multilayer structure of SATURN. When the two layers are in contact with each other, charges are induced in the copper and the PTFE due to triboelectrification (Figure 4a). PTFE, which has a greater electron affinity, is able to gain electrons from the copper and becomes negatively charged, whereas the copper layer on the paper becomes positively charged. Subsequent separation of the paper and the PTFE (Figure 4b) induces a potential difference across the two copper electrodes, causing current to flow from the paper toward the PTFE when the device is connected to an external load. This flow of current reverses the polarity (Figure 4c) of charges on the two copper electrodes (i.e., now the copper on PTFE has more positive charge than the copper layer on the paper). The next compression results in the paper moving toward the PTFE again, resulting in a reversed direction of current flow (Figure 4d), completing the cycle of electricity generation.

3. PERFORMANCE CHARACTERIZATION

3.1. Self-sustainable microphone

Even though SATURN is self-sustainable, flexible, and thin, its quality as a sensor is comparable to an active microphone that consumes power. After structural optimization of SATURN, the best acoustic sensitivity of –25.63 dB (relative mV/Pa) is achieved at 1000 Hz. The resulting SATURN patch has a circular shape with a 16 cm^2 area, a grid pattern of holes 0.4 mm in diameter with 0.2 mm spacing, and glue attachments of the two layers at the center and at eight equally distant points around the edges of the PTFE (Figure 5). In this configuration, the SATURN microphone compares favorably to an active microphone (MEMS ADMP-401 and iPhone INMP441) for frequencies as high as 5000 Hz (Figure

Figure 2. Fabrication process: (1) preparation of micro-hole paper; (2) deposition of copper layer; (3) attaching copper tape as electrodes; (4) stacking paper and PTFE; and (5) gluing paper and PTFE. All dimensions are in mm.

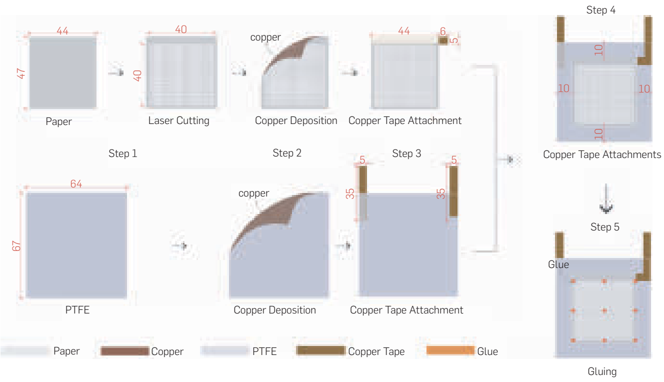


Figure 3. SATURN's structural design parameters.

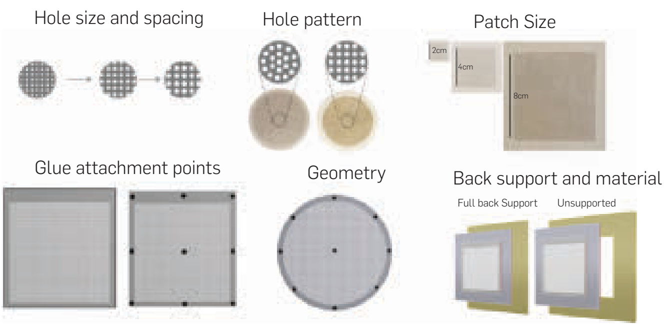


Figure 4. Cycle of electricity generation process under external acoustic excitation.

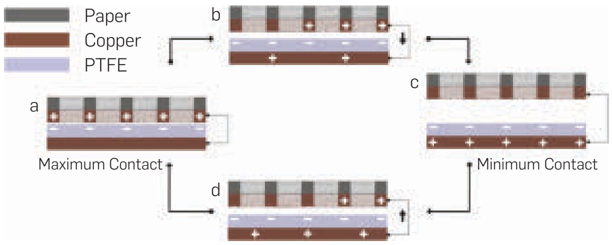


Figure 5. Sensitivity across acoustic bands (20 Hz to 20 kHz).

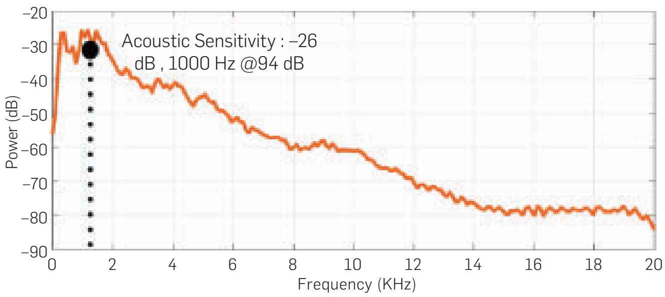
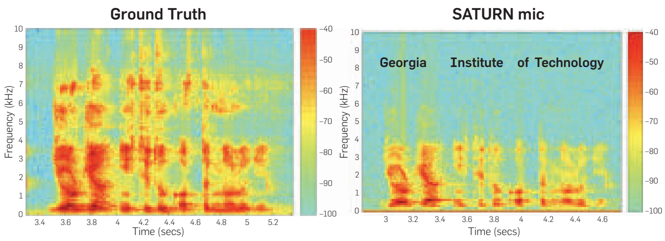


Figure 6. Acoustic sensitivity of SATURN sensor compared to an active microphone.



6). Because approximately 90% of the acoustic information for human speech lies within this range,[4] SATURN can be used a good quality microphone for a variety of applications.

Bending of SATURN reduces the acoustic sensitivity as the bend angle increases due to the increase in the stiffness of the structure, which results in lesser vibration of the layers. At a bending angle of 45°, SATURN is still a usable microphone and is comparable to an active microphone up to 3000 kHz, allowing capture of more than 60% of the sound information of speech.[4] See Figure 7.

3.2. Loud sound energy harvester

Next we look at SATURN as an energy harvester for loud acoustic events. We performed an experiment to determine the peak voltage and peak power of the SATURN microphone at its resonant frequency as functions of an external load resistance. We analyzed the 4 × 4 cm^2 SATURN microphone patch as a power harvester when exposed to loud sounds (100 dB). The voltage is approximately 0.5 V_{pp} at 150 Hz and rises to a maximum at 2.5 V_{pp} at 250 Hz and then falls below 1.0 V_{pp} at 350 Hz. The same behavior is shown in the power curve, with a maximum power of 6499 nW (Figure 8 and Table 1).

4. SELF-SUSTAINABLE COMPUTATIONAL SYSTEM DESIGN

To use SATURN in a practical application, it needs to be embedded in a self-sustaining computational system. In this section, we look at three different ways of building such system using SATURN. The first system leverages SATURN as a self-sustainable microphone to transfer a wide-spectrum audio sounds (e.g., speech or taps). The second and the third system demonstrate how we can leverage SATURN as a sound energy harvester to build a self-sustainable event detection systems for loud sounds. These self-sustainable wireless sensing and communication systems maintain the form factor of the computational material although still being potentially cheap to manufacture.

Figure 7. Experimental results for the effect of flexibility: change in acoustic sensitivity (1000 Hz at 94 dB obtained for different radii of curvature).

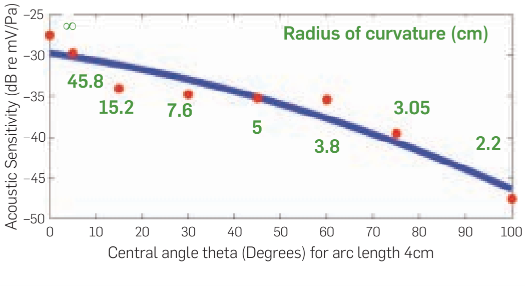
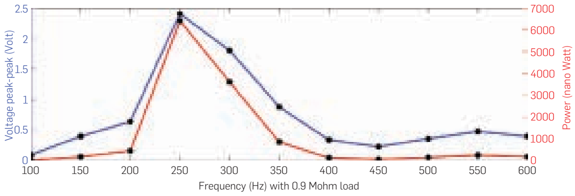


Figure 8. Voltage and power generated at different working frequencies.



4.1. Analog backscatter-based communication for sound and vibration

We combine the self-sustainable mechanical vibration sensing property of SATURN with a self-sustainable communication technique called analog backscatter, to build a thin flexible material tag that can self-sustainably both sense and communicate audio. The word backscatter in analog backscatter means reflection. We can explain analog backscatter by using a simple analogy of reflection of light rays (Figure 9). When light hits a mirror, it bounces off the surface at an angle equal to the angle of the incoming light wave. However, when light hits an irregular surface it gets diffused, resulting in modulation in both the intensity and angle of reflection of the light waves. Analysis of the received diffused light can help deconstruct the shape of the irregular surface from which it is reflected. Similar to the light waves, when radio frequency (RF) waves are incident on a specially designed tag, they get modulated in amplitude and phase. This reflected RF from the tag can be processed at the receiver to extract information of the phenomenon which caused the modulation. In our tag design, the modulation is due to voltage generated in SATURN resulting from sound or mechanical vibration (Figure 10).

Our prototype tag consists of SATURN, a junction gate field-effect transistor (JFET) that acts as a voltage-controlled impedance device, and a printed antenna (Figure 11).[1] Sounds and vibrations in the environment result in generation of voltage in SATURN which in turn changes the impedance in the circuit via the JFET. This change in impedance changes the radar cross section of the antenna resulting in amplitude modulation of the incoming RF waves. Our tag is an example of a computational material for audio and

Figure 9. RF analog backscatter can be explained using an analogy to the modulation of intensity and angle of reflection of light from a rough surface.

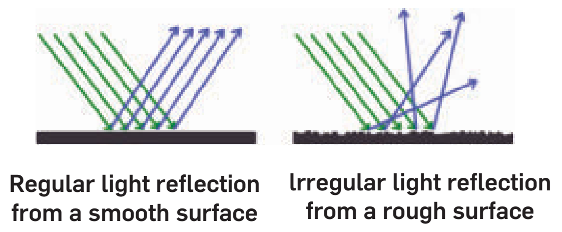
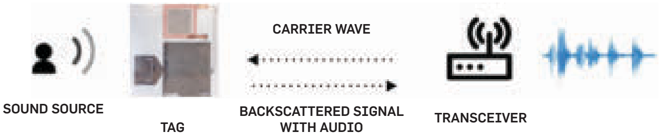


Figure 10. Self-sustainable sound sensing and communication architecture using SATURN and analog backscatter technique.



vibration sensing and communication. The simple circuit design of our tag maintains the thin, flexible form factor of SATURN and can easily be embedded in everyday objects and surfaces.

Amplitude modulation-based analog backscatter does not allow unique IDs for different SATURN patches, but other methods such as frequency shift keying (FSK)[11] can allow this capability by the addition of a subhundred microwatt power energy harvesting method based on easily available sources such as wireless energy or light sources in the room.

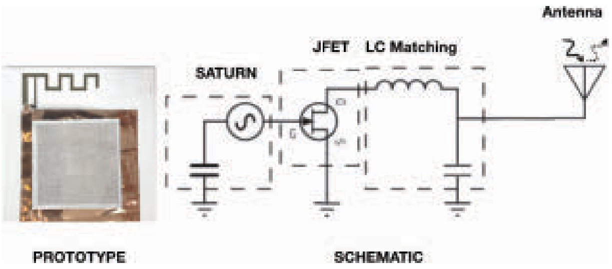
4.2. Flip-bit data storage and RF interrogation system

The 6.9 μW of power generated by SATURN due to a loud sound can be used to flip a bit in nonvolatile memory to record the occurrence. Considering the maximum power transfer theorem (Jacobi's law), the usable power we can obtain is approximately 50%. Thus, we might harvest up to 3.4 μJ/s. The energy required to program a "1" into NAND flash memory is 2 μJ.[8] Given that the sounds we wish to monitor will probably last for several seconds, there is more than enough power to record the acoustic event. Going further, SRAM bits can be flipped at approximately 10–100 pW of power,[5, 10] suggesting that rudimentary computation might be performed to determine if the flash memory bit should be written. Recorded bits might be read later using a passive RFID interrogation system, which can both read the recorded state and reset it.

4.3. Ultralow power radio communication system

SATURN could power longer range radio transmitters allowing real-time alerts to sounds that exceed a loudness threshold. Talla et al.[13] have recently demonstrated a 915 MHz analog LoRa backscatter communications device that can communicate at greater than 11 bits/s although hundreds of meters away from its RF source and receiving antenna. Although their system currently uses a battery, their theoretical IC design consumes only 9.25 μW of power. With sound events lasting on the order of seconds, one can imagine a SATURN-based system storing power until it has enough to enable a 915 MHz backscatter transmission to the receiving antenna, announcing the event. As long as the event continues to occur, the SATURN system can transmit alerts every few seconds to a remote monitoring station. In this manner, acoustic environmental monitoring can be performed without the cost and environmental difficulties of batteries.

Figure 11. Tag consisting of simple circuitry—SATURN sensor connected to printed antenna using a JFET and an impedance matching circuit.



5. APPLICATION SCENARIOS

In this section, we explore how self-sustainable computational systems based on SATURN can be used in everyday scenarios.

5.1. Ubiquitous microphone for interaction and control

The thin and flexible form factor of SATURN allows it to be placed on many different surfaces (Figure 12). SATURN patches might be placed on walls or lamp shades in the home to act as a baby monitor or extend the range of audio input for home assistants (e.g., Amazon Echo or Google Home). In addition to audio sensing, SATURN can be used as a vibration sensor to detect simple input tap touch to control objects. For example, imagine a SATURN patch in the form factor of a Post-it Note which could be placed on walls or tables as a wireless light switch.

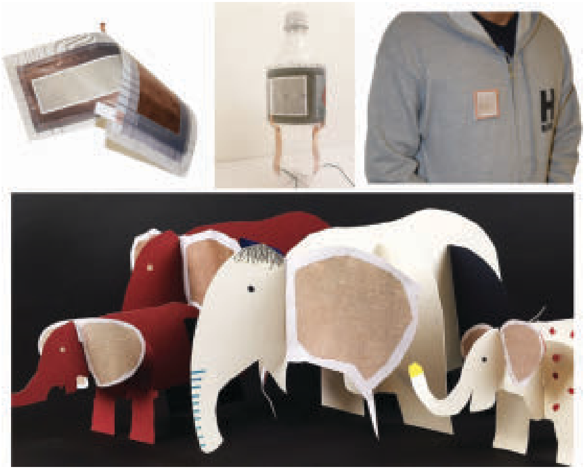
5.2. Audio sensing Post-it Notes for infrastructure mapping and authentication

Imagine placing a SATURN Post-it Note at a conference room door entrance which only authorized users can open using a password consisting of unique combinations of blow, whistle, or tap (Figure 13). In a public restroom scenario, SATURN Post-it Notes can be placed on restroom doors; a special sequence of tapping can start a maintenance request. Multiple SATURN patches with pre-mapped IDs can be placed at multiple places in a nursing home and could be used for easy access to an emergency help button by tapping or speaking to the patch. The main advantage of such audible Post-it Notes is that they are cheap and disposable, reducing concerns of them being lost or stolen.

5.3. Context sensing and localization

Using multiple SATURN patches can allow beam-forming such that individual voices may be better isolated

Figure 12. SATURN is flexible and can be made in different shapes and sizes, allowing instrumentation of everyday objects such as a soda bottle, shirt, and paper crafts.

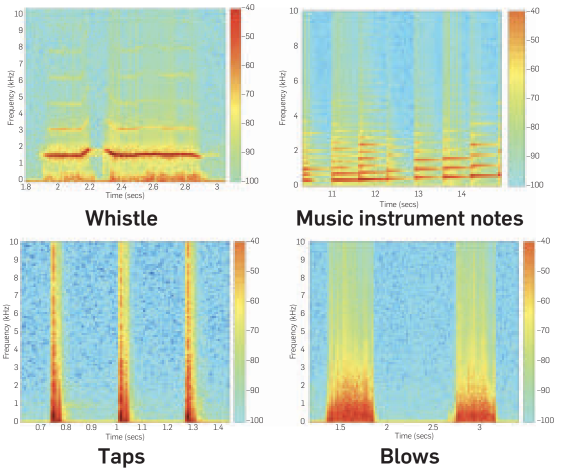


acoustically. For example, placing several SATURN patches on a conference table can help localize a speaker to improve speech quality (Figure 14). Outdoor arrays of SATURN patches might sense loud sounds, such as a gunshot, and transmit the sound to indoor transceivers for real-time triangulation (Figure 15).

5.4. Industrial and home monitoring

In more industrial environments, SATURN can be used for acoustic failure monitoring and diagnostics in places where it is difficult or dangerous for humans to access. For example, SATURN patches might be placed on turbines in a nuclear power plant to monitor them for vibrations that indicate damage. Arranging SATURN patches in an array[2] increases the range of audio sensing to almost 10 m allowing monitoring of large rooms. Thus, a SATURN array can be used as a remote glass-break detector for security systems in office buildings.

Figure 13. Examples of sounds sensed by SATURN.



Whistle

Music instrument notes

Taps

Blows

Figure 14. Multiple SATURN patches placed on a table can localize speakers.

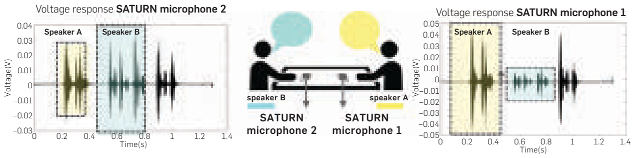
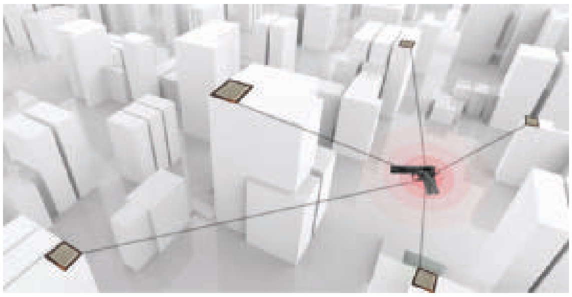


Figure 15. Outdoor SATURN patches can sense and communicate gunshot sounds to indoor transceivers for real-time triangulation.



5.5. Ambient monitoring of noise pollution

Imagine an airport located in the center of a city, such as San Jose International in California, which would like to monitor its acoustic environment so as to not exceed safe noise levels for its employees and to keep overly loud aircraft noise footprints within airport boundaries (Figure 16). A SATURN-based system can be tiled on various buildings and at various distances on the runway. As planes take off, they generate loud sounds due to engines, fans, and air turbulence. The peak in the sound spectrum generated by aircraft is near the 200–300 Hz band (Figure 45 in Khorrami et al.[6]) with decibel levels reaching 105 dB_{SPL} at 5 m. These values are consistent with the resonant frequency of the SATURN patch and would result in generation of power >6.9 μW accumulated over different frequency bands. This power could be used to communicate the event by either of the two self-sustainable systems described in Sections 4.2 and 4.3.

Other work zones that might monitor for sound thresholds exceeding human hearing tolerance include construction zones, mines, music venues, power stations, airports, space-ports, and military environments. Similarly, SATURN-based sensors might be used for monitoring catastrophic events such as landslides, avalanches, polar ice calving, mine caveins, and mine gas explosions.

5.6. Localization of tanks in war zones

In a more futuristic application, the United Nations might drop SATURN-based sensors from an airplane into a conflict zone. The sensors would monitor the acoustic and ground vibration environment for the movement of tanks, heavy chemical transports, mortars, or exploding ordnance. As tanks go by, the sound flips a bit in the sensor. Later, an official with a reading device might sweep the field to interrogate the sensors (Figure 17). In a more extreme scenario,

Figure 16. Recording a loud acoustic event using power generated from a SATURN microphone.

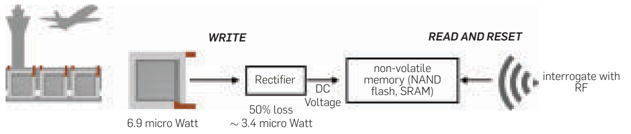
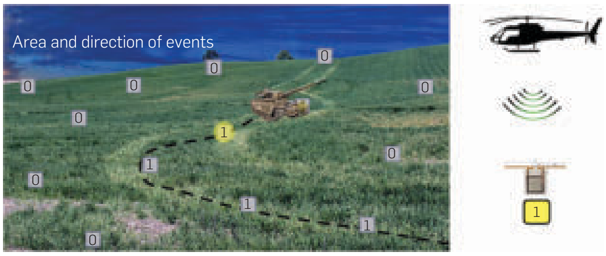


Figure 17. Dropping remotely interrogable SATURN sensors to monitor tanks and ordnance in a war zone.



a low flying drone aircraft might sweep a strong RF signal over the region and record which sensors report hearing an event. The pattern of reporting sensors can reveal the direction of travel of a vehicle and point to possible hiding areas for that equipment, providing proof for treaty violations.

6. EMERGING RESEARCH THEMES

Work on SATURN suggests several research directions for computational materials in the future. We describe some of those themes here.

6.1. Self-sustaining sensing opportunities

There are many other opportunities for designing self-sustaining sensors based on different energy harvesting phenomena and for placing them in the context of self-sustainable computational systems. SATURN is just one example; we can design many other TENG-based self-powered sensors in different form factors for varied applications. A photodiode or solar cell can be re-cast as a self-sustainable motion sensor whose rate of wireless transmission is directly coupled to the amount of light to which the sensor is exposed. Another example is to re-think batteries, which generate power due to an electrochemical reaction, as sensors. Design changes to the cathode, anode, and electrolyte can make the cell relatively inert until exposed to a catalyst such as air, water, or other chemicals, thus making the cell a self-powered sensor for detecting (and communicating) the presence of the chemical catalyst. A third opportunity regards thermoelectric harvesters, which generate energy due to heat flow and temperature gradients. Imagine a steam pipe constructed out of the dissimilar metals typically used for thermoelectric generators. When steam flows through the pipe, the temperature difference would cause energy to be generated which would power the wireless transmission of packets. The higher the temperature difference, the more the power and the higher the rate of packet transmissions.

6.2. Transitioning from devices to materials

Although such self-sustainable sensors can be made as individual macro units, current technology trends support a lower level integration of the sensing into the material itself. Advancements in printed and flexible electronics are enabling the production of self-sustainable sensors in thin and flexible form factors that can be conveniently added to current materials. New research in flexible antennas, transistors, and integrated circuits[9] demonstrates how simple computation and communication can be added although maintaining flexibility and low cost. Finally, with radio backscatter technology and applications improving rapidly,[11, 20] a surprisingly low amount of energy needs to be generated locally for communication to the external infrastructure.

6.3. Rethinking traditional semiconductor manufacturing techniques

In an ideal situation, we would like these paper microphones to be cheap and disposable so we would not worry if they are lost or stolen. Currently, the bill of materials for a single Post-it Note-sized SATURN microphone is less than a cent, but its manufacturing cost is still high due to the way we are depositing copper on paper and PTFE. When SATURN is placed in a self-sustainable computational scenario, the cost is driven higher depending on the active transistor component being used. This expense suggests reexamining the manufacturing process such that it can be more efficiently scaled. How should the traditional semiconductor industry best support large-scale ubiquitous sensing? How can we change traditional manufacturing techniques to be able to support applications where objects and surfaces have computation embedded in it.

6.4. Designing the user experience

Computational materials promise to further blur the divide between the physical and digital worlds, opening a very interesting set of research challenges for HCI and design practitioners. How do we design user experiences around this new technology? What tools and design frameworks can be adapted for an Internet of Materials? When the form factor of computing is more like the objects we use in crafting, how does that change the way we think about designing user experiences? Indeed, as the playful art object in Figure 12 suggests, computation that looks more like paper or other material inspires more creative uses by those familiar with those materials. And computation that looks and feels more like everyday objects will change the way we as humans experience, understand, and build relationships with that technology. It also creates opportunities and challenges for infusing computation with values, such as sustainability, through the deliberate choice of materials used to create a computational effect.

6.5. In-material privacy frameworks

One of the biggest challenges and opportunities with ubiquitous computing systems is designing for privacy.[7, 19] Privacy-aware designs often focus on both technological and social approaches.[7, 12] What new approaches will the Internet of Materials era inspire? What privacy models will need changing? One opportunity comes in the form of the privacy principles of choice and consent and proximity and locality.[7] Computational materials focus on local sensing, which may provide users a natural mental model of their range of operation. For example, SATURN microphone patches could be constructed at a physical level to require tapping with a finger to prime the backscatter circuit before audio could be transmitted (e.g., when a speaker taps a microphone to ask "is this mic on?"). Furthermore, the patch can be constructed to have a limited sensing range for human voice.

Another way to design SATURN for privacy is to tune the resonant frequencies of the patch to focus on only certain types of sound or vibration. A SATURN patch for monitoring an industrial machine might only listen to certain low-pitched hums, outside normal speech ranges, which indicate upcoming failures. Similarly, a glass-break detector might be tuned for high pitches outside of human hearing ranges. It is our hope that by focusing on embedding "privacy in the material structure" as a first-class research priority, we will also find ways of improving the function of the device along the often-related dimensions of power usage, networking, and user interface.

6.6. Building a community of multidisciplinary researchers

Building SATURN involves solving technical, system level, and design challenges that span many fields. Materials science is required to design SATURN patches; mechanical engineering helps characterize the effect of vibration on a patch; wireless, low-power electronics is necessary for building self-sustainable communication; flexible electronics are required for manufacturing the prototype; and design and HCI knowledge help us explore applications in everyday settings. Creating devices for an Internet of Materials needs researchers who can adopt an aggressively multidisciplinary mindset to collaborate and learn the language of many different fields.

7. CONCLUSION

SATURN is an initial example of a computational material. It senses vibration and can convey that information wirelessly without batteries or wired power sources. Its simple multilayer construction resembles an everyday material; it looks and feels like paper, yet it behaves like a wireless microphone. The multidisciplinary, out-of-the-box thinking that resulted in SATURN's creation and application is meant to inspire a new direction for computing. Using power (requiring no external power source), cost (large-scale manufacturing with simple materials), and form factor (looking more like an everyday object) as driving factors leads to a rethinking of ubiquitous computing that can drive the community forward for the coming decades in the same way that Weiser's vision propelled us for the past few decades.

Acknowledgments

We would like to thank our collaborators at Georgia Tech's Nanotechnology Lab—Prof. Zhong Lin Wang, Yi-Cheng Wang, Steven L. Zhang, and Zhengjun Wang—for their work and guidance on the original version of the SATURN paper.[3] We thank Diego Osorio and Fereshteh Shahmiri for their help in making diagrams for the original SATURN paper, some of which have been adopted in this review paper. We thank Mohit Gupta for the device vibration simulations, and we are grateful to Jin Yu and Hyunjoo Oh for their help in making the teaser image for this paper.

References

1. Arora, N., Abowd, G.D. ZEUSSS: Zero energy ubiquitous sound sensing surface leveraging triboelectric nanogenerator and analog backscatter communication. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings* (2018), ACM, 81–83.
2. Arora, N., Xue, Q., Bansal, D., McAughan, P., Bahr, R., Osorio, D., Ma, X., Sample, A.P., Starner, T.E., Abowd, G.D. Surface++: A scalable and self-sustainable wireless sound sensing surface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services* (2019), ACM, 543–544.
3. Arora, N., Zhang, S.L., Shahmiri, F., Osorio, D., Wang, Y.-C., Gupta, M., Wang, Z., Starner, T., Wang, Z.L., Abowd, G.D. SATURN: A thin and flexible self-powered microphone leveraging triboelectric nanogenerator. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*, 2 (2018), 60.
4. Galbrun, L., Kitapci, K. Speech intelligibility of English, Polish, Arabic and Mandarin under different room acoustic conditions. *Appl. Acoust.* 114 (2016), 79–91.
5. Gilbert, N., Zhang, Y., Dinh, J., Calhoun, B., Hollmer, S. A 0.6 v 8 pJ/write non-volatile CBRAM macro embedded in a body sensor node for ultra low energy applications. In *Proceedings of the 2013 Symposium on VLSI Circuits* (*VLSIC*, 2013), IEEE, 204–205.
6. Khorrami, M.R., Fares, E., Casalino, D. Towards full aircraft airframe noise prediction: Lattice Boltzmann simulations. In *20th AIAA/CEAS Aeroacoustics Conference* (2014), 2481.
7. Langheinrich, M. Privacy by design—Principles of privacy-aware ubiquitous systems. In *International Conference on Ubiquitous Computing* (2001), Springer, 273–291.
8. Mohan, V., Bunker, T., Grupp, L., Gurumurthi, S., Stan, M.R., Swanson, S. Modeling power consumption of NAND flash memories using flashpower. *IEEE Trans. Comp. Aid. Des. Integ. Circuits Sys. 7*, 32 (2013), 1031–1044.
9. Myny, K. The development of flexible integrated circuits based on thin-film transistors. *Nat. Electr. 1*, 1 (2018), 30–39.
10. Raikwal, P., Neema, V., Verma, A. High speed 8t SRAM cell design with improved read stability at 180 nm technology. In *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA, volume 2)* (2017), IEEE, 563–568.
11. Ranganathan, V., Gupta, S., Lester, J., Smith, J.R., Tan, D. RF Bandaid: A fully-analog and passive wireless interface for wearable sensors. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (volume 2(2))* (2018), 79.
12. Richards, N.M. The dangers of surveillance. *Harvard Law Rev. 7*, 126 (2013), 1934–1965.
13. Talla, V., Hessar, M., Kellogg, B., Najafi, A., Smith, J.R., Gollakota, S. LoRa backscatter: Enabling the vision of ubiquitous connectivity. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (volume 1(3))* (2017), 105.
14. Wang, Z.L. Triboelectric nanogenerators as new energy technology and self-powered sensors—Principles, problems and perspectives. *Faraday Discuss.* 176 (2015), 447–458.
15. Wang, Z.L. On Maxwell's displacement current for energy and sensors: The origin of nanogenerators. *Mater. Today 2*, 20 (2017), 74–82.
16. Wang, Z.L., Chen, J., Lin, L. Progress in triboelectric nanogenerators as a new energy technology and self-powered sensors. *Energy Environ. Sci 8*, 8 (2015), 2250–2282.
17. Wang, Z.L., Wang, A.C. Triboelectric nanogenerator for self-powered flexible electronics and internet of things. In *Meeting Abstracts (26)* (2018), The Electrochemical Society, 1533–1533.
18. Weiser, M. The computer for the 21st century. *Scientific American 3*, 265 (1991), 94–105.
19. Weiser, M. Some computer science issues in ubiquitous computing. *Commun. ACM 7*, 36 (1993), 75–84.
20. Zhang, Y., Iravantchi, Y., Jin, H., Kumar, S., Harrison, C. Sozu: Self-powered radio tags for building-scale activity sensing. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), ACM, 973–985.
21. Zou, H., Zhang, Y., Guo, L., Wang, P., He, X., Dai, G., Zheng, H., Chen, C., Wang, A.C., Xu, C., et al. Quantifying the triboelectric series. *Nat. Commun. 1*, 10 (2019), 1427.

Nivedita Arora, Thad Starner, and Gregory D. Abowd ({nivedita.arora,thad,arbowd}@gatech.edu), School of Interactive Computing, Georgia Institute of Technology, USA.

CAREERS

Boston College

Tenure Track Assistant Professor of Computer Science

The Computer Science Department of Boston College seeks a tenure-track Assistant Professor beginning in the 2021-2022 academic year. Successful candidates for the position will be expected to develop strong research programs that can attract external funding in an environment that also values high-quality undergraduate teaching. Outstanding candidates in all areas of Computer Science will be considered, with a preference for those who demonstrate a potential to contribute to cross-disciplinary teaching and research in conjunction with the planned Schiller Institute for Integrated Science and Society at Boston College.

A Ph.D in Computer Science or a closely related discipline is required. See cs.bc.edu and https://www.bc.edu/bc-web/centers/schiller-institute.html for more information. Application review is ongoing.

Applicants should submit a cover letter, a detailed CV, and teaching and research statements. Arrange for three confidential letters of recommendation to be uploaded directly to Interfolio. To apply go to: https://apply.interfolio.com/79609.

Boston College conducts background checks as part of the hiring process. Information about the University and our department is available at bc.edu and cs.bc.edu.

Boston College is a Jesuit, Catholic university that strives to integrate research excellence with a foundational commitment to formative liberal arts education. We encourage applications from candidates who are committed to fostering a diverse and inclusive academic community. Boston College is an Affirmative Action/Equal Opportunity Employer and does not discriminate on the basis of any legally protected category including disability and protected veteran status. To learn more about how BC supports diversity and inclusion throughout the university, please visit the Office for Institutional Diversity at http://www.bc.edu/offices/diversity.

Department of Electrical and Computer Engineering
Graduate School of Engineering and Management
Air Force Institute of Technology (AFIT)
Dayton, Ohio

Faculty Position

The Department of Electrical and Computer Engineering at the Air Force Institute of Technology is seeking applications for a tenured or tenure-track faculty position. All academic ranks will be considered. Applicants must have an earned doctorate in Electrical Engineering, Computer Engineering, Computer Science, or a closely affiliated discipline by the time of their appointment (anticipated 1 September 2021).

We are particularly interested in applicants specializing in one or more of the following areas: autonomy, artificial intelligence / machine learning, navigation with or without GPS, cyber security, and VLSI. Candidates in other areas of specialization are also encouraged to apply. This position requires teaching at the graduate level as well as establishing and sustaining a strong DoD relevant externally funded research program with a sustainable record of related peer-reviewed publications.

The Air Force Institute of Technology (AFIT) is the premier Department of Defense (DoD) institution for graduate education in science, technology, engineering, and management, and has a Carnegie Classification as a High Research Activity Doctoral University. The Department of Electrical and Computer Engineering offers accredited M.S. and Ph.D. degree programs in Electrical Engineering, Computer Engineering, and Computer Science as well as an MS degree program in Cyber Operations.

Applicants must be U.S. citizens. Full details on the position, the department, applicant qualifications, and application procedures can be found at *http://www.afit.edu/ENG/*. Review of applications will begin on 4 January 2021. The United States Air Force is an equal opportunity, affirmative action employer.

Harvard John A. Paulson School of Engineering and Applied Sciences

Tenure-Track Faculty Positions in Computer Science and Theory of Quantum Information and Computation

The Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS) seeks applicants for a tenure-track position in Computer Science at the assistant or associate professor level. Additionally, the SEAS area of Computer Science and the Harvard Quantum Initiative seeks applicants for a tenure-track position in the area of the theory of quantum information and/or computation at the assistant or associate professor level. Both have an expected start date of July 1, 2021.

For the position in Computer Science, we invite applications in all areas of Computer Science. Areas of special interest include (but are not limited to) both machine learning and algorithms (broadly construed). For the position in the area of the theory of quantum information and/or computation, areas of interest may include but are not limited to quantum algorithms, communication, complexity, control, cryptography, and information processing.

Computer Science at Harvard benefits from outstanding undergraduate and graduate students, world-leading faculty, significant industrial collaboration, and substantial support from SEAS. Starting in Spring 2021, Computer Science will be among the founding occupants of Harvard's new Science and Engineering Complex, with ample space for growth and surrounded by state-of-the-art facilities. Information about Harvard's current faculty, research, and educational programs in computer science is available at http://www.seas.harvard.edu/computer-science.

The associated Institute for Applied Computational Science (http://iacs.seas.harvard.edu), Data Science Initiative (https://datascience.harvard.edu/), and Center for Research on Computation and Society (http://crcs.seas.harvard.edu/) foster connections among computer science, applied math, data science, and various domain sciences at Harvard through its graduate programs and events.

We seek candidates who have a strong research record and a commitment to undergraduate and graduate teaching and training. We particularly encourage applications from historically underrepresented groups, including women and minorities. A doctorate or terminal degree in a related field is required by the expected start date for each position.

Required application documents include a cover letter, CV, a statement of research interests, a teaching statement, and up to three representative papers. In addition, we ask for a statement describing efforts to encourage diversity, inclusion, and belonging, including past, current, and anticipated future contributions in these areas. Candidates are also required to submit the names and contact information for at least three and up to five references, and the application is complete only when three letters have been submitted. At least one letter must come from someone who has not served as the candidate's undergraduate, graduate, or postdoctoral advisor. We encourage candidates to apply by December 15, 2020, but will continue to review applications until the position is filled. Applicants for the tenure-track position in Computer Science can apply online at http://academicpositions.harvard.edu/postings/9834. Applicants for the tenure-track position in the area of the theory of quantum information and/or computation can apply online at http://academicpositions.harvard.edu/postings/9835.

We are an equal opportunity employer, and all qualified applicants will receive consideration for employment without regard to race, color, religion, sex, national origin, disability status, protected veteran status, gender identity, sexual orientation, pregnancy and pregnancy-related conditions or any other characteristic protected by law.

Rutgers University

Tenure Track Assistant Professor in Theoretical Computer Science

The Computer Science Department at Rutgers University invites applications for a Tenure-Track Assistant Professor position in Theoretical Computer Science. We welcome candidates working on computational complexity theory but outstanding applicants in all areas of TCS will be considered. Consistent with the aims of the Simons Junior Faculty Fellows program, which provides partial funding, the department also welcomes applicants who are most affected by the COVID-19 pandemic: postdocs and new PhDs.

The appointment will start September 1, 2021. Responsibilities include research in the area of Theoretical Computer Science, supervision of PhD students, and teaching undergraduate and graduate level courses in Computer Science. Pursuit of external research funding is expected.

Qualifications: Successful completion of a PhD or equivalent in Computer Science or a closely related field is required by the start date.

Application Instructions: Applicants should submit their CV, a research statement addressing both past and future work, a teaching statement, and contact information for three references at http://jobs.rutgers.edu/postings/120527. Applications received by January 15, 2021 will be given priority.

For questions, contact: martin@farach-colton.com.

The CS Department is strongly committed to increasing the diversity of our faculty and welcomes applications from women, dual-career couples, historically underrepresented populations and candidates with disabilities. Offer is contingent upon successful completion of all pre-employment screenings. Rutgers University is an affirmative action/equal opportunity employer.

Southern University of Science and Technology (SUSTech)

Faculty Positions in Computer Science and Engineering

The Department of Computer Science and Engineering (CSE, http://cse.sustc.edu.cn/en/), Southern University of Science and Technology (SUSTech) has multiple Tenure-track faculty openings at all ranks. We are looking for outstanding candidates with demonstrated research achievements and keen interest in teaching, in the following areas (but are not limited to):

- ▸ Data Science
- ▸ Artificial Intelligence
- ▸ Computer Systems
- ▸ Software Engineering (senior positions only)
- ▸ Cognitive Robotics and Autonomous Systems
- ▸ Programming Languages and Compilers

Applicants should have an earned Ph.D. degree and demonstrated achievements in both research and teaching. The teaching language at SUSTech is bilingual, either English or Putonghua. It is perfectly acceptable to use English in all lectures, assignments, exams. In fact, our existing faculty members include several non-Chinese speaking professors.

The Department of Computer Science and Engineering at SUSTech was founded in 2016. It has 27 tenured or tenure-track professors, all of whom hold doctoral degrees or have years of experience in overseas universities. 24 of the 27 were recruited from outside the mainland China. Among them, three are IEEE fellows, two were editors-in-chief of IEEE journals. The department

EPFL

Faculty Position in Computer and Communication Sciences

at the Ecole polytechnique fédérale de Lausanne (EPFL)

The School of Computer and Communication Sciences (IC) at EPFL invites applications for tenure-track faculty positions in **all** areas of computer and communication sciences. Some areas of particular interest this year include unconventional computing (e.g., applied quantum computing, DNA computing), programming languages and verification, and intelligent systems.

Senior faculty appointments may be possible.

We seek candidates with an outstanding academic record and a strong commitment to teaching and mentoring students.

EPFL offers its faculty excellent students from all over the world, competitive salaries, generous research support, and outstanding research infrastructure. Switzerland has an exceptionally high human development index and is consistently ranked top in economic competitiveness and innovation.

To apply, follow the application procedure at

https://facultyrecruiting.epfl.ch/position/23691281

You will be required to submit in PDF form a cover letter, a curriculum vitae including a publication list, brief statements of research and teaching interests, and contact information (name, postal address, and email) of 3 references for junior positions and at least 5 for senior positions

Screening will start on **December 1, 2020**. Further questions can be addressed to:

Prof. George Candea

Chair of the Faculty Recruiting Committee

School of Computer and Communication Sciences

ic_erecruiting@epfl.ch

For additional information on EPFL and IC, please visit: www.epfl.ch or https://ic.epfl.ch

EPFL is an equal opportunity employer and family friendly university. It is committed to increasing the diversity of its faculty. It strongly encourages women to apply.

is expected to grow to 50 tenured or tenure-track faculty members eventually, in addition to teaching-only professors and research-only professors. In 2019, the department has secured external grants of RMB123.5 million (approx USD18 million).

SUSTech is a pioneer in higher education reform in China. The mission of the University is to become a globally recognized research university which emphasizes academic excellence and promotes innovation, creativity and entrepreneurship. Set on five hundred acres of wooded landscape in the picturesque Nanshan (South Mountain) area, the campus offers an ideal environment for learning and research.

SUSTech is committed to increase the diversity of its faculty, and has a range of family-friendly policies in place. The university offers competitive salaries and fringe benefits including medical insurance, retirement and housing subsidies, which are among the best in China. Salary and rank will commensurate with qualifications and experience. More information can be found at http://talent.sustech.edu.cn/.

We provide some of the best start-up packages in the sector to our faculty members, including one PhD studentship per year, in addition to a significant amount of start-up grant (which can be used to fund additional PhD students and postdocs, research travels, and research equipment).

To apply, please provide a cover letter identifying the primary areas of your research and listing your five best publications, curriculum vitae, and research and teaching statements, and forward them to cshire@sustech.edu.cn.

University of Michigan College of Engineering

Multiple Tenure-Track and Teaching Faculty (Lecturer) Positions

Computer Science and Engineering (CSE) at the University of Michigan College of Engineering invites applications for multiple tenure-track and teaching faculty (lecturer) positions, as part of its aggressive long-term growth plan. We seek exceptional candidates in all areas across computer science and computer engineering, with special emphasis on candidates at the early stages of their careers; we also have a targeted search for an endowed professorship in theoretical computer science (the Fischer Chair).

Qualifications include an outstanding academic record; an awarded or expected doctorate (or equivalent) in computer science, computer engineering, or a related area; and a strong commitment to teaching and research. We seek faculty members who commit to excellence in graduate and undergraduate education, will develop impactful, productive and novel research programs, and will contribute to the department's goal of eliminating systemic racism and sexism by embracing our culture of Diversity, Equity and Inclusion.

We encourage candidates to apply as soon as possible. Positions remain open until filled and applications can be submitted throughout the year. For more details on these positions and to apply, please visit https://cse.engin.umich.edu/about/faculty-hiring/.

The University of Michigan is one of the world's leading research universities, consisting of highly ranked departments and colleges across engineering, sciences, medicine, law, business, and the arts, with a commitment to interdisciplinary collaboration. CSE is a vibrant and innovative community, with over 70 world-class faculty members, over 300 graduate students, and a large and illustrious network of alumni. Ann Arbor is known as one of the best small cities in the nation, and the University has a strong dual-career assistance program.

Michigan Engineering's vision is to be the world's preeminent college of engineering serving the common good. This global outlook, leadership focus, and service commitment permeate our culture. Our vision is supported by our mission and values that, together, provide the framework for all that we do. Information about our vision, mission and values can be found at http://strategicvision.engin.umich.edu/.

The University of Michigan has a storied legacy of commitment to Diversity, Equity and Inclusion (DEI). The Michigan Engineering component of the University's comprehensive, five-year, DEI strategic plan—with updates on our programs and resources dedicated to ensuring a welcoming, fair, and inclusive environment—can be found at: http://www.engin.umich.edu/college/about/diversity.

The University of Michigan is an equal opportunity/affirmative action employer and is responsive to the needs of dual-career families.

Penn Engineering

MULTIPLE FACULTY POSITIONS

Department of Electrical and Systems Engineering

The School of Engineering and Applied Science at the University of Pennsylvania is growing its faculty by 33% over a five-year period. As part of this initiative, the **Department of Electrical and Systems Engineering** is engaged in an aggressive, multi-year hiring effort for multiple tenure-track positions at all levels. Candidates must hold a Ph.D. in Electrical Engineering, Computer Engineering, Systems Engineering, or related area. The department seeks individuals with exceptional promise for, or proven record of, research achievement, who will take a position of international leadership in defining their field of study and who will excel in undergraduate and graduate education. Leadership in cross-disciplinary and multi-disciplinary collaborations is of particular interest. We are interested in candidates in all areas that enhance our research strengths in:

1. **Nanodevices and nanosystems** (nanoelectronics, MEMS/NEMS, power electronics, nanophotonics, nanomagnetics, quantum devices, integrated devices and systems at nanoscale);
2. **Circuits and computer engineering** (analog, RF, mm-wave, digital circuits, emerging circuit design, computer engineering, IoT, beyond 5G, and cyber-physical systems);
3. **Information and decision systems** (control, optimization, robotics, data science, machine learning, communications, networking, information theory, signal processing).

Prospective candidates in all areas are strongly encouraged to address large-scale societal problems in energy, transportation, health, agriculture, food and water, economic and financial networks, social networks, critical infrastructure, and national security. We are especially interested in candidates whose interests are aligned with the school's strategic plan, *https://www.seas.upenn.edu/about/penn-engineering-2020/*

Diversity candidates are strongly encouraged to apply. Interested persons should submit an online application at *https://www.ese.upenn.edu/faculty-staff/* and include curriculum vitae, research, teaching, and diversity statements, and at least three references. Review of applications will begin on January 4, 2021.

The University of Pennsylvania values diversity and seeks talented students, faculty and staff from diverse backgrounds. The University of Pennsylvania is an equal opportunity and affirmative action employer. Candidates are considered for employment without regard to race, color, sex, sexual orientation, gender identity, religion, creed, national or ethnic origin, citizenship status, age, disability, veteran status or any other legally protected class. Questions or concerns about this should be directed to the Executive Director of the Office of Affirmative Action and Equal Opportunity Programs, University of Pennsylvania, 421 Franklin Building, 3451 Walnut Street, Philadelphia, PA 19104-6205; or (215) 898-6993 (Voice) or (215) 898-7803 (TDD).

[CONTINUED FROM P. 104] extra minus sign in the program, purely by accident, so we adopted counterclockwise as our norm. We made a few thousand bolts with the same program, then realized they would not work in ordinary nuts, so we updated the nut-milling program. Soon we built our clocks with number circle and hand movements going the other way, and began using our left arms when shaking human hands. As the British can testify, driving cars on the proverbially wrong side of the road is exactly as good, and distinguishes their culture from the rest of the world."

Frederic went on to explain all very technical skills were programmed into the machinery, so human workers could shift from job to job, gaining new experiences and avoiding drudgery. For example, each month, one member of the community would be selected to make the regular deliveries of milk and ice, and would temporarily take the name Hickman, derived poetically from Eugene O'Neill's play, *The Iceman Cometh*, about the human need for extreme hope. Burrhus asked why ice needed to be delivered, and Frederic explained, "We have iceboxes in our homes, rather than electric refrigerators. Every winter, our robots cut the ice off the top of our lake, filling the ice barn and covering with sawdust to prevent the huge pile from melting during the warm months. The Hickman also delivers milk and other liquids in big reusable bottles, as was common a century ago."

They then talked about the economy, and Burrhus learned Walden Three had an internal virtual currency called Goldmarks. The name came from the inventor of the long-playing phonograph record, Peter Goldmark, who in a 1972 government report and *Scientific American* article had predicted something like the Internet would render cities irrelevant, and allow people to live happily again in small towns. Goldmarks essentially were the same as hours in time banking, paying people equally, and totally separated from the dollars of the surrounding economy. Because residents of Walden Three earned no dollars, they paid no income tax.

Walden Three had an internal virtual currency the Goldmark, named for Peter Goldmark, who predicted in 1972 that something like the Internet would render cities irrelevent.

Frederic had been smiling and lecturing his brother excitedly, but then he became gloomy. "Burr, our community is really in trouble right now, because the outside governments want to impose real estate, income, and sales taxes on us, as they greedily are changing the laws concerning volunteering and barter organizations. I don't know how we can survive. The great communes of the 19th century all produced salable agriculture. Oneida made animal traps, silk garments, and eventually became a silverware company. The Shakers made furniture and beautiful wooden boxes. The Amana communities began manufacturing refrigerators, of all things! But we have nothing to sell." Burrhus expressed sympathy, but tended to think the governments were right, given they protected the commune and provided a range of implicit services and social insurance.

At suppertime they sat at a picnic table, surrounded by all members of Walden Three, as they sang, "'Tis the gift to be simple, 'tis the gift to be free… in the valley of love and delight." Frederic explained that was the Shaker theme song, especially relevant because his community was struggling to decide its policy about family relationships. The Shakers never divided into couples and produced no children. Then a second song began: "We will build us a dome on our beautiful plantation, and we'll all have one home, and one family relation." That was the theme song of Oneida, where all members were married to all others, and many children were born according to spiritual eugenic rules imposed by its supreme leader. Burrhus decided he definitely did not want his own personal relationships decided by his crazy brother, or even a democratic vote of all members of this group of losers. However, he shuddered in realization that he also was a loser, because the Mars colonization effort had failed.

He thought about the other 22 survivors, a number that could be prime factored into 2 times 11, or 11 married couples. One of them he personally cared very much about, Catniss, might be his partner if she joined Walden Three with him. Rather than strengthening their relationship, living together in the tiny spaceship crew compartment for months had driven them apart. Surviving on the hostile Martian surface had required all 41 colonists to function like gears in a mechanical system, rather than as human beings sharing personal relationships. They returned to Earth exhausted in every way.

Rather than reunite with Catniss, Burrhus wondered if he should recruit some of the Mars survivors to help Walden Three, whether or not they joined it. Perhaps starting with one would be a proper experiment, and he immediately thought of three from which to select the initial research subject:

1. Catniss Rockefeller had been the chief of Mars logistics and was trained in economics; she could find a way to turn big profits trading the Goldmark currency to pay the taxes, or marketing unconventional picnic tables.
2. The French international coordinator for the expedition, Paris Mason, had a law degree, so potentially he could delay the taxes indefinitely through constant litigation in court.
3. Communication programmer Thomas Sanderson could put his hacker skills to a new use: threatening the external economy by erasing bank accounts to make governments relent.

Which one should he recruit to this challenging cause, if any?

Then Burrhus noted that the numbers 1, 2, 3 were all primes, and there was another prime in their sequence that was more realistic: zero.

William Sims Bainbridge (wsbainbridge@hotmail.com) is a sociologist who taught classes on crime and deviant behavior at respectable universities before morphing into a computer scientist, editing an encyclopedia of human-computer interaction, writing many books on things computational, from neural nets to virtual worlds to personality capture, then repenting and writing harmless fiction.

From the intersection of computational science and technological speculation, with boundaries limited only by our ability to imagine what could be.

DOI:10.1145/3427948

William Sims Bainbridge

Future Tense
Walden Three

Can humanity take its next step forward by taking a step back?

WHEN NAVIGATION PROGRAMMER Burrhus Skinner returned from NASA's aborted three-year attempt to colonize the planet Mars, he faced a difficult transition. The plan had been to send a fleet of spaceships carrying supplies, equipment, and 41 people to establish the beginning of a socially and biologically perfect society. Everyone agreed to the same set of social norms, and humans would be the only living species on the red planet. Of course there could be no birds, because the atmosphere was too thin for flying, but also no mammals, reptiles, bacteria or viruses, so infectious diseases would be impossible. But equipment kept breaking down, in the worst case depriving an outpost with 13 unlucky people of electric power; five did not survive the crash of a transport, and the prime number who returned to Earth was 23. For several weeks they waited in quarantine, as germs were gently returned to their digestive systems so they could survive the biological complexity of their home world.

This gave Burrhus time to ponder what he would do next. He had not abandoned his utopian ambitions, so he naturally thought of the Walden Three experimental community set up by his crazy brother, Frederic Skinner. In preparation for visiting, he read the 1948 novel *Walden Two* written by their psychologist ancestor, B. F. Skinner. Then he read the 1854 memoire, *Walden*, by Henry David Thoreau, that was the origin of the community's name. *Walden* had emphasized the human experience of unity with nature during social isolation. *Walden Two* had argued that a group of like-minded people could voluntarily decide upon a strict set of norms, then use behavioral psychology to ensure that all members of the group followed them. In 1974, Kathleen Kinkade published *A Walden Two Experiment* about a very real community she helped found in 1967 that continues to thrive.

On the basis of its website, given Burrhus had not talked with Frederic for many years, he gathered the principle added by Walden Three was distributed manufacturing, in which AI-controlled 3D printing and milling machines could help workshops in the community produce most of the products it needed.

When Burrhus arrived at Walden Three, he found a few decrepit old houses and a set of four new log cabins arranged in the shape of a plus sign around some picnic tables where workers could socialize over communal meals. In his habitually over-emotional voice, Frederic explained, "We cut the wood parts of these tables and all other furniture in the pine log workshop to the East, and mill metal components like the screws to hold the parts of a table together were automatically milled in the workshop to the West. Come, I was about to assemble one, and you can do it to see how much you would enjoy joining us!"

The many wooden boards for a picnic table were already shaped and drilled for assembly with screws, but Burrhus was amazed to see that each screw was as fat as his thumb. Frederic explained that the computer milling machine in the opposite log cabin could not handle iron or steel, so they made all metal components by recycling copper pipes and aluminum trash, which required screws to be big to be strong enough. Burrhus complained, "Oh, look Fred, these screws your robot miller made don't work! Using my screwdriver, I turn and turn, and they don't go in!"

"Oh, Burr, sorry about that miscommunication," Frederic said. "The screws work fine if you turn them counterclockwise rather than clockwise. There was an

[CONTINUED ON P. 103]

IMAGE BY ANDRIJ BORYS ASSOCIATES, USING SHUTTERSTOCK

IPDPS 2021 Portland, OR

Portland, Oregon USA 17-21 May 2021

ipdps.org

35th IEEE International Parallel and Distributed Processing Symposium

ANNOUNCING 19 IPDPS 2021 WORKSHOPS

IPDPS Workshops are the "bookends" to the three-day conference technical program of contributed papers, invited speakers, student programs, and industry participation. They provide the IPDPS community an opportunity to explore special topics and present work that is more preliminary or cutting-edge than the more mature research presented in the main symposium. Each workshop has its own website and submission requirements, and the submission deadline for most workshops is after the main conference author notification dates.

IPDPS WORKSHOPS MONDAY 17 May 2021

HCW	Heterogeneity in Computing Workshop
RAW	Reconfigurable Architectures Workshop
HiCOMB	High Performance Computational Biology
GrAPL	Graphs, Architectures, Programming, and Learning
EduPar	NSF/TCPP Workshop on Parallel and Distributed Computing Education
HIPS	High-level Parallel Programming Models and Supportive Environments
AsHES	Accelerators and Hybrid Emerging Systems
PDCO	Parallel / Distributed Combinatorics and Optimization
APDCM	Advances in Parallel and Distributed Computational Models

IPDPS WORKSHOPS FRIDAY 21 MAY 2021

JSSPP	Job Scheduling Strategies for Parallel Processing
PDSEC	Parallel and Distributed Scientific and Engineering Computing
iWAPT	Automatic Performance Tuning
MPP	Parallel Programming Models - Emerging Technologies on Machine Learning Acceleration
SNACS	Scalable Networks for Advanced Computing Systems
PAISE	Parallel AI and Systems for the Edge
RADR	Resource Arbitration for Dynamic Runtimes
ScaDL	Scalable Deep Learning over Parallel And Distributed Infrastructures
HPS	High-Performance Storage
ParSocial	Parallel and Distributed Processing for Computational Social Systems

Sponsored by

IEEE COMPUTER SOCIETY TCPP Technical Committee on Parallel Processing

acm In-Cooperation

ACM SIGARCH sighpc

GENERAL CO-CHAIRS

David Bader (New Jersey Institute of Technology, USA)
Aparna Chandramowlishwaran (University of California, Irvine, USA)

PROGRAM CHAIR

Karen Karavanic (Portland State University, USA)

WORKSHOPS CHAIR and VICE-CHAIR

Erik Saule (University of North Carolina Charlotte, USA)
Jaroslaw Zola (The State University of New York at Buffalo, USA)

STUDENT PROGRAM

The conference plans to hold the traditional PhD forum of poster presentations, to provide mentoring in scientific writing and presentation skills, and to create opportunities for students to hear from and interact with senior researchers attending the conference.

INDUSTRY PARTICIPATION

There are several ways for industry to partner with IPDPS and share the benefits of associating with our international community of top researchers and practitioners in fields related to parallel processing and distributed computing. IPDPS is a "walk-up-and-talk" venue that encourages industry partners to use the conference as a way to introduce their technology in an informal setting.

IMPORTANT DATES

Conference Preliminary Author Notification	*Dec. 8, 2020*
Conference Final Author Notification	*Jan. 19, 2021*
Workshops' Call for Papers Deadlines	*Most Fall in Late Dec. and Jan.*

IPDPS 2021 VENUE & PROGRAM

Portland, Oregon sits on the Columbia and Willamette rivers in the shadow of snow-capped Mount Hood and is known for its parks, roses and rhododendrons, bridges and bicycle paths, as well as for its eco-friendliness. Planning for the 35th edition of IPDPS will be informed by experience in 2020 and collaboration with the IPDPS community to continue the programming that promotes international cooperation in seeking to apply computer science technology to the betterment of our global village. Whether presented in person or virtually, every paper accepted for IPDPS 2021 will be published in the proceedings, and valuable interaction whatever the platform will be assured.

Trust Extension as a Mechanism for Secure Code Execution on Commodity Computers
Bryan Jeffrey Parno
Text Data Management and Analysis
A Practical Introduction to Information Retrieval and Text Mining
ChengXiang Zhai
Sean Massung
The Sparse Fourier Transform
Theory and Practice
Haitham Hassanieh
Frontiers of Multimedia Research
Shih-Fu Chang
Editor
Smarter than Their Machines
Oral Histories of Pioneers in Interactive Computing
John Cullinane
The VR Book
Human-Centered Design for Virtual Reality
Jason Jerald, Ph.D.
Verified Functional Programming in Agda
Aaron Stump
A Framework for Scientific Discovery through Video Games
Seth Cooper
An Architecture for Fast and General Data Processing on Large Clusters
Matei Zaharia
The Handbook of Multimodal-Multisensor Interfaces
Volume 1
Foundations, User Modeling, and Common Modality Combinations