

# COMMUNICATIONS

CACM.ACM.ORG OF THE ACM 11/2021 VOL.64 NO.11

## Special Section on China Region



MIP\* = RE

There Is No AI Without Data

Privacy Engineering Superheroes

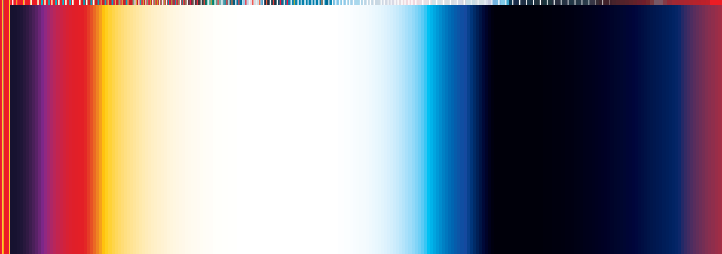
Holograms on the Horizon?



# SIGGRAPH ASIA 2021 TOKYO

CONFERENCE 14 - 17 DECEMBER 2021  
EXHIBITION 15 - 17 DECEMBER 2021  
TOKYO INTERNATIONAL FORUM, JAPAN

[sa2021.siggraph.org](http://sa2021.siggraph.org)



ONE GIFT

# *WORLD-CHANGING IMPACT*

25 MILLION  
SMALL STEPS TOWARD

ADVANCING KNOWLEDGE  
REVOLUTIONIZING EDUCATION  
TRANSFORMING SOCIETY

115+  
FACULTY

1,700+  
UNDERGRADUATE  
STUDENTS

1,000+  
GRADUATE  
STUDENTS

Learn about the Elmore family's contribution at:

[PURDUE.LINK/ELMORE](https://www.purdue.edu/link/elmore)



Elmore Family School of Electrical  
and Computer Engineering

## Departments

- 5 **Vardi's Insights**  
**The Paradox of Choice in Computing-Research Conferences**  
*By Moshe Y. Vardi*
- 
- 7 **Career Paths in Computing**  
**Computing Enabled Me To ... Grace Hopper, Minicomputers, and Megabytes: It's a Fun Career**  
*By Ann Moffatt*
- 
- 8 **BLOG@CACM**  
**Assessing Internet Software Engineering, Encouraging Competitions**  
Andrei Sukhov considers the potential for reducing international tensions through competitive events, while Vivek S. Buzruk looks at the evolution of teaching Internet software engineering.
- 
- 139 **Careers**

## Last Byte

- 144 **Future Tense**  
**World of Hackcraft**  
An obsessive gamer's quest for the absolutely most significant computer ever.  
*By William Sims Bainbridge*

## News



- 11 **Qubit Devices Inch Toward Reality**  
Key questions and challenges remain, including how to scale qubit devices while reducing noise and errors to the point where the devices become useful.  
*By Samuel Greengard*
- 
- 14 **Holograms on the Horizon?**  
Machine learning drives toward 3D imaging on the move.  
*By Chris Edwards*
- 
- 17 **Filtering for Beauty**  
Social media "influencers" use augmented reality filtering apps to appear more beautiful, together, and cool. Results may vary.  
*By Esther Shein*

## Viewpoints

- 20 **Legally Speaking**  
**Text and Data Mining of In-Copyright Works: Is It Legal?**  
How copyright law might be an impediment to text and data mining research.  
*By Pamela Samuelson*
- 
- 23 **Privacy**  
**Privacy Engineering Superheroes**  
Privacy engineers are essential to both preventing and responding to organizational privacy problems.  
*By Lea Kissner and Lorrie Cranor*
- 
- 26 **Computing Ethics**  
**Shaping Ethical Computing Cultures**  
Lessons from the recent past.  
*By Katie Shilton, Megan Finn, and Quinn DuPont*
- 
- 30 **Education**  
**Explicative Programming**  
Making Computational Thinking relevant to schools.  
*By Alexander Repenning and Ashok Basawapatna*
- 
- 34 **Viewpoint**  
**Medical Artificial Intelligence: The European Legal Perspective**  
Although the European Commission proposed new legislation for the use of "high-risk artificial intelligence" earlier this year, the existing European fundamental rights framework already provides some clear guidance on the use of medical AI.  
*By Karl Stöger, David Schneeberger, and Andreas Holzinger*
- 
- 37 **Viewpoint**  
**We Are Not Users: Gaining Control Over New Technologies**  
Seeking a more selective approach to technology usage.  
*By Yoram Reich and Eswaran Subrahmanian*

Special Section



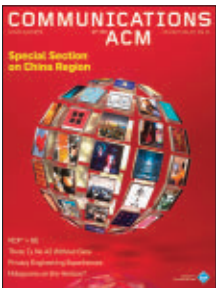
40

**40 China Region Special Section**  
In this issue, we return our editorial spotlight to China, with a special section that presents new and emerging technologies and educational reforms that have taken place since our first look at the region in November 2018.

Practice

**94 What Went Wrong?**  
Why we need an IT accident investigation board.  
*By Poul-Henning Kamp*

**Q** Articles' development led by **acmqueue**  
queue.acm.org



**About the Cover:**  
A kaleidoscope of imagery from the China section (p. 40) shows a rich selection of topics ranging from start-ups and digital commerce, to AI enterprises and education, to accessibility enhancements, to blockchain and supercomputing advancements, and much more.  
Cover illustration by Spooky Pooka at Debut Art.

IMAGES IN COVER COLLAGE: Coffee bin photo courtesy of SaturnBird/Labbrand.com; Students photo courtesy of Tsinghua University; Processor photo courtesy of China Education and Research Network (CERNET); Cloud Ginger image courtesy of CloudMinds Technology Inc; Taobao app photo courtesy of Alibaba Group/alzila.com; Smart China expo and robot cafe photos by helloabe/Shutterstock.com; Subway photo by Tada Images/Shutterstock.com; Exhibit photo by ItzaVU/Shutterstock.com; Xpeng store photo by ItzaVU/Shutterstock.com. Additional stock images from Shutterstock.com.

Contributed Articles



98

**98 There Is No AI Without Data**  
Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises.  
*By Christoph Gröger*



Watch the author discuss this work in the exclusive *Communications* video.  
<https://cacm.acm.org/videos/no-ai-without-data>

**109 Inferring and Improving Street Maps with Data-Driven Automation**  
Automatic map inference, data refinement, and machine-assisted map editing promises more accurate map datasets.  
*By F. Bastani, S. He, S. Jagwani, E. Park, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and M.A. Sadeghi*

Research Highlights

- 120 Technical Perspective**  
**Finding the Sweet Spot Amid Accuracy and Performance**  
*By Pascal Van Hentenryck*
- 
- 121 Multi-Itinerary Optimization as Cloud Service**  
*By Alexandru Cristian, Luke Marshall, Mihai Negrea, Flavius Stoichescu, Peiwei Cao, and Ishai Menache*
- 
- 130 Technical Perspective**  
**On Proofs, Entanglement, and Games**  
*By Dorit Aharonov and Michael Chapman*
- 
- 131 MIP\* = RE**  
*By Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen*



ACM, the world's largest educational and scientific computing society, delivers resources that advance computing as a science and profession. ACM provides the computing field's premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.

**Executive Director and CEO**

- Vicki L. Hanson
- Deputy Executive Director and COO**  
Patricia Ryan
- Director, Office of Information Systems**  
Wayne Graves
- Director, Office of Financial Services**  
James Schembari
- Director, Office of SIG Services**  
Donna Cappel
- Director, Office of Publications**  
Scott E. Delman

**ACM COUNCIL**

- President**  
Gabriele Kotsis
- Vice-President**  
Joan Feigenbaum
- Secretary/Treasurer**  
Elisa Bertino
- Past President**  
Cherri M. Pancake
- Chair, SGB Board**  
Jeff Jortner
- Co-Chairs, Publications Board**  
Joseph Konstan and Divesh Srivastava
- Members-at-Large**  
Nancy M. Amato; Tom Crick;  
Susan Dumais; Mehran Sahami;  
Alejandro Saucedo
- SGB Council Representatives**  
Sarita Adve and Jeanna Neefe Matthews

**BOARD CHAIRS**

- Education Board**  
Elizabeth Hawthorne and Chris Stephenson
- Practitioners Board**  
Terry Coatta

**REGIONAL COUNCIL CHAIRS**

- ACM Europe Council**  
Chris Hankin
- ACM India Council**  
Abhiram Ranade
- ACM China Council**  
Wenguang Chen

**PUBLICATIONS BOARD**

- Co-Chairs**  
Joseph Konstan and Divesh Srivastava
- Board Members**  
Jonathan Aldrich; Jack Davidson;  
Chris Hankin; Mike Heroux;  
James Larus; Marc Najork;  
Holly Rushmeier; Eugene H. Spafford;  
Bhavani Thuraisingham

**DIGITAL LIBRARY BOARD**

- Chair**  
Jack Davidson
- Board Members**  
Phoebe Ayers; Yannis Ioannidis;  
Michael Ley; Michael L. Nelson;  
Louisa Raschid; Theo Schlossnagle;  
Julie Williamson

# COMMUNICATIONS OF THE ACM

Trusted insights for computing's leading professionals.

*Communications of the ACM* is the leading monthly print and online magazine for the computing and information technology fields. *Communications* is recognized as the most trusted and knowledgeable source of industry information for today's computing professional. *Communications* brings its readership in-depth coverage of emerging areas of computer science, new trends in information technology, and practical applications. Industry leaders use *Communications* as a platform to present and debate various technology implications, public policies, engineering challenges, and market trends. The prestige and unmatched reputation that *Communications of the ACM* enjoys today is built upon a 50-year commitment to high-quality editorial content and a steadfast dedication to advancing the arts, sciences, and applications of information technology.

**STAFF**

**DIRECTOR OF PUBLICATIONS**

Scott E. Delman  
cacm-publisher@cacm.acm.org

**Executive Editor**

Diane Crawford  
**Managing Editor**  
Thomas E. Lambert

**Senior Editor**

Ralph Raiola  
**Senior Editor/News**  
Lawrence M. Fisher

**Web Editor**

David Roman  
**Editorial Assistant**  
Danbi Yu

**Art Director**

Andrij Borys  
**Associate Art Director**  
Margaret Gray

**Assistant Art Director**

Mia Angelica Balaquiot  
**Production Manager**  
Bernadette Shade  
**Intellectual Property Rights Coordinator**  
Barbara Ryan  
**Advertising Sales Account Manager**  
Ilia Rodriguez

**Columnists**

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

**CONTACT POINTS**

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

**REGIONAL SPECIAL SECTIONS**

**Co-Chairs**  
Jakob Rehof, Haibo Chen, and P J Narayanan  
**Board Members**  
Sherif G. Aly; Panagiotis Fatourou;  
Chris Hankin; Sue Moon; Tao Xie;  
Kenjiro Taura; David Padua

**WEBSITE**

http://cacm.acm.org

**WEB BOARD**

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

**AUTHOR GUIDELINES**

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

**ACM U.S. TECHNOLOGY POLICY OFFICE**

Adam Eisgrau  
Director of Global Policy and Public Affairs  
1701 Pennsylvania Ave NW, Suite 200,  
Washington, DC 20006 USA  
T (202) 580-6555; acmpo@acm.org

**COMPUTER SCIENCE TEACHERS ASSOCIATION**

Jake Baskin  
Executive Director

**EDITORIAL BOARD**

**EDITOR-IN-CHIEF**

Andrew A. Chien  
aacm@cacm.acm.org

**Deputy to the Editor-in-Chief**

Morgan Denlow  
cacm.deputy.to.eic@gmail.com

**SENIOR EDITOR**

Moshe Y. Vardi

**NEWS**

**Co-Chairs**

Marc Snir and Tom Conte

**Board Members**

Mei Kobayashi; Rajeev Rastogi;  
François Sillion

**VIEWPOINTS**

**Co-Chairs**

Tim Finin; Susanne E. Hambrusch;  
John Leslie King

**Board Members**

Virgilio Almeida; Terry Benzel; Michael L. Best;  
Judith Bishop; Lorrie Cranor;  
James Grimmelmann; Mark Guzdial;  
Haym B. Hirsch; Anupam Joshi; Richard Ladner;  
Carl Landwehr; Beng Chin Ooi; Francesca Rossi;  
Len Shustek; Loren Terveen;  
Marshall Van Alstyne; Susan J. Winter

**Q PRACTICE**

**Co-Chairs**

Stephen Bourne and Theo Schlossnagle

**Board Members**

Eric Allman; Samy Bahra; Peter Bailis;  
Betsy Beyer; Terry Coatta; Stuart Feldman;  
Nicole Forsgren; Camille Fournier;  
Jessie Frazelle; Benjamin Fried; Tom Killalea;  
Tom Limoncelli; Kate Matsudaira;  
Marshall Kirk McKusick; Erik Meijer;  
George Neville-Neil; Jim Waldo;  
Meredith Whittaker

**CONTRIBUTED ARTICLES**

**Co-Chairs**

James Larus and Gail Murphy

**Board Members**

Robert Austin; Nathan Baker; Kim Bruce;  
Alan Bundy; Peter Buneman;  
Premkumar T. Devanbu; Jane Cleland-Huang;  
Yannis Ioannidis; Rebecca Isaacs;  
Trent Jaeger; Somesh Jha; Gal A. Kaminka;  
Ben C. Lee; Igor Markov; m.c. schraefel;  
Hannes Werthner; Ryan White;  
Reinhard Wilhelm; Rich Wolski

**RESEARCH HIGHLIGHTS**

**Co-Chairs**

Shriram Krishnamurthi  
and Orna Kupferman

**Board Members**

Martin Abadi; Amr El Abbadi;  
Animashree Anandkumar; Sanjeev Arora;  
Michael Backes; Maria-Florina Balcan;  
Azer Bestavros; David Brooks; Stuart K. Card;  
Jon Crowcroft; Lieven Eeckhout;  
Alexei Efros; Bryan Ford; Alon Halevy;  
Gernot Heiser; Takeo Igarashi;  
Srinivasan Keshav; Sven Koening;  
Ran Libeskind-Hadas; Karen Liu;  
Joanna McGrenere; Tim Roughgarden;  
Guy Steele, Jr.; Robert Williamson;  
Margaret H. Wright; Nikolai Zeldovich;  
Andreas Zeller

**Association for Computing Machinery (ACM)**

1601 Broadway, 10<sup>th</sup> Floor  
New York, NY 10019-7434 USA  
T (212) 869-7440; F (212) 869-0481

**ACM Copyright Notice**

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

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

**Subscriptions**

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

**Single Copies**

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

**ACM ADVERTISING DEPARTMENT**

1601 Broadway, 10<sup>th</sup> Floor  
New York, NY 10019-7434 USA  
T (212) 626-0686  
F (212) 869-0481

**Advertising Sales Account Manager**

Ilia Rodriguez  
ilia.rodriguez@hq.acm.org

**Media Kit** acmm mediasales@cacm.org

**COMMUNICATIONS OF THE ACM**

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

**POSTMASTER**

Please send address changes to *Communications of the ACM* 1601 Broadway, 10<sup>th</sup> Floor New York, NY 10019-7434 USA

Printed in the USA.



Association for Computing Machinery





Moshe Y. Vardi

DOI:10.1145/3488554

# The Paradox of Choice in Computing-Research Conferences

**I**N THE EARLY 1970s, after a major struggle with Soviet authorities who denied Soviet Jews permission to emigrate, primarily to Israel, such permission was granted on a limited basis. Small waves of Soviet emigrants landed in Israel, making a transition from the Soviet economy to Israel's Western economy. I remember stories about their encounter with Israeli supermarkets. In the Soviet Union, a friend told me, if you wanted to buy canned peas, there was one available type of canned peas. In an Israeli supermarket, there were a dozen different types of canned peas. The choice was so difficult that in many cases ex-Soviet shoppers just gave up and left the store without making a purchase.

This phenomenon was studied by American psychologist Barry Schwartz in his 2004 book, *The Paradox of Choice—Why More Is Less*. “Autonomy and freedom of choice are critical to our well-being, and choice is critical to freedom and autonomy” wrote Schwartz. “Nonetheless, though modern Americans have more choice than any group of people ever has before, and thus, presumably, more freedom and autonomy, we don't seem to be benefiting from it psychologically.”

I believe the problem of over-choice also describes what is happening today with computing-research conferences. When COVID-19 erupted in early 2020, many conferences had to pivot from an in-person mode to a fully remote mode. By April 2020, ACM had released guidelines for “Best Practices for Virtual Conferences.”<sup>a</sup> Computing-research conferences also quickly discovered the

standard conference registration fee of say, around US\$500–\$700, is not realistic for virtual conferences. Registration fees have been reduced drastically. Many conferences are free to non-authors or have a nominal fee. I now can sit in my home office and “attend” dozens of conferences per year, at almost no cost.

While the characters in David Lodge's 1984 novel, *Small World: An Academic Romance*, spend their time traveling from conference to conference, I can hop from conference to conference at any time, all from the convenience of my home. Yet I do not, as I have too many choices. In addition to conferences, events of various types are taking place all over the world, and most are open. The conference/event world is truly flat now. Dozens of email notices land in my inbox every week. But processing this stream of notices is by itself a time-consuming task, so most of these notices languish unopened. I must confess that the extent of my conference participation has shrunk over the past year and a half, rather than expanded. Too much choice.

It may seem this problem has been created by the pandemic, and it will go away when the pandemic ends, hopefully sooner rather than later, but I do not think so. Undoubtedly, we are all eager to resume social contact, and there will be a rush to attend in-person conferences as soon as it is safe to do so. But there is a growing realization that COVID-19 may have been just a dress rehearsal for a much larger crisis—the climate crisis. The slew of extreme-weather events over the past few years has demonstrated compellingly that the issue is not about a rise in planetary temperatures in the year 2100, but with a climate that is getting more extreme *right now!*

In my January 2020 *Communications*

column, “Publish *and* Perish,”<sup>b</sup> I asked: “Can we reduce the carbon footprint of computing-research publishing?” In a March 2020 *Communications* Viewpoint, “Conferences in an Era of Expensive Carbon,”<sup>c</sup> Pierce et al. make several specific recommendations for ACM to reduce the carbon footprint of its conferences. But I believe the proposals in both columns do not go far enough. The very idea that each paper publication must involve conference travel is not morally acceptable anymore. Virtual and hybrid conferences are here to stay, I am convinced, which means the paradox of choice is here to stay.

The fundamental problem, I believe, is the current computing-research publication system conflates research publishing with community building. Other disciplines view the two as two distinct activities of their research community. In the past I have pointed out the weaknesses of the conference system as a vehicle for scholarly publishing, but conferences were effective community-building vehicles. Virtual conferences have yet to become effective community-building vehicles. We may need in-person conferences for community building, but not so many!

In my opinion, the status quo is simply not acceptable anymore. Change is imperative. *We must change!* ■

<sup>b</sup> <https://cacm.acm.org/magazines/2020/1/241717-publish-and-perish/fulltext/>

<sup>c</sup> <https://cacm.acm.org/magazines/2020/3/243024-conferences-in-an-era-of-expensive-carbon/fulltext>

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

Copyright held by author.

<sup>a</sup> <https://www.acm.org/media-center/2020/april/virtual-conferences-best-practices/>

# SHAPE THE FUTURE OF COMPUTING. JOIN ACM TODAY.

[www.acm.org/join/CAPP](http://www.acm.org/join/CAPP)

## SELECT ONE MEMBERSHIP OPTION

### ACM PROFESSIONAL MEMBERSHIP:

- Professional Membership: \$99 USD
- Professional Membership plus ACM Digital Library: \$198 USD (\$99 dues + \$99 DL)

### ACM STUDENT MEMBERSHIP:

- Student Membership: \$19 USD
- Student Membership plus ACM Digital Library: \$42 USD
- Student Membership plus Print *CACM* Magazine: \$42 USD
- Student Membership with ACM Digital Library plus Print *CACM* Magazine: \$62 USD

- Join ACM-W:** ACM-W supports, celebrates, and advocates internationally for the full engagement of women in computing. Membership in ACM-W is open to all ACM members and is free of charge.

## PAYMENT INFORMATION

Name

Mailing Address

City/State/Province

ZIP/Postal Code/Country

- Please do not release my postal address to third parties

Email Address

- Yes, please send me ACM Announcements via email
- No, please do not send me ACM Announcements via email

- AMEX  VISA/MasterCard  Check/money order

Credit Card #

Exp. Date

Signature

### Purposes of ACM

ACM is dedicated to:

- 1) Advancing the art, science, engineering, and application of information technology
- 2) Fostering the open interchange of information to serve both professionals and the public
- 3) Promoting the highest professional and ethics standards

By joining ACM, I agree to abide by ACM's Code of Ethics ([www.acm.org/code-of-ethics](http://www.acm.org/code-of-ethics)) and ACM's Policy Against Harassment ([www.acm.org/about-acm/policy-against-harassment](http://www.acm.org/about-acm/policy-against-harassment)).

I acknowledge ACM's Policy Against Harassment and agree that behavior such as the following will constitute grounds for actions against me:

- Abusive action directed at an individual, such as threats, intimidation, or bullying
- Racism, homophobia, or other behavior that discriminates against a group or class of people
- Sexual harassment of any kind, such as unwelcome sexual advances or words/actions of a sexual nature

## BE CREATIVE. STAY CONNECTED. KEEP INVENTING.



ACM General Post Office  
P.O. Box 30777  
New York, NY 10087-0777

1-800-342-6626 (US & Canada)  
1-212-626-0500 (Global)  
Hours: 8:30AM - 4:30PM (US EST)

Fax: 212-944-1318  
[acmhelp@acm.org](mailto:acmhelp@acm.org)  
[www.acm.org/join/CAPP](http://www.acm.org/join/CAPP)





# CAREER PATHS IN COMPUTING

DOI:10.1145/3485446

Computing enabled me to . . .

## Grace Hopper, Minicomputers, and Megabytes: It's a Fun Career



### NAME

**Ann Moffatt**

### BACKGROUND

**Born in 1939 in London, England  
migrated to Australia in 1974**

### CURRENT JOB TITLE/EMPLOYER

**Retired**

### EDUCATION

**Grad Dip Technology  
Management, Macquarie Univ.  
Honorary Doctorate  
awarded in 2006, Univ. of  
Southern Queensland**

**M**Y CAREER STARTED when I joined Kodak in the U.K. in 1959, where I was taught to program by Conway Berners-Lee, father of Sir Tim Berners-Lee, the WWW inventor. At that time, we only knew of about 300 stored program computers in the world, although there were probably 300 more in 'secret' places like the military or government.

By 1963, computing in the U.K. surpassed the rest of the world. The British government decided to make the world's most powerful computer: the mighty Atlas, the first computer with an operating system. The manufacturer, Ferranti, couldn't get it to work and

asked would-be buyers to send their best programmers to help. Kodak sent me.

The sales price was about £UK 3 million (about \$150 million USD today). The sales manager told his team if Ferranti sold three Atlases to the Russian government it would solve Russian computer needs to the year 2000. It was the time of MAD (Mutually Assured Destruction), when computers guided nuclear weapons and supported the space race to get a man to the moon. Atlas was less powerful than the MAC I'm using today.

As Special Ambassador for Univac, Grace Hopper, the inventor of COBOL (the most widely used programming language in the 1960s to 1980s), toured the world giving lectures about the industry. I had the pleasure of escorting her to various British Computer Society functions whenever she visited the U.K.

After a dinner in 1973, she asked if we would like to see the new computer Univac had loaned her. She dived into her handbag and brought out an object the size of a cigarette packet. We all stared, amazed, as she opened the box and picked up an even smaller object. Grace proceeded to tell us the impossibly small computer had a 64-kilobyte COBOL compiler. We wanted to see it in action, so someone brought over a teletype with a printer, and from the side of the device Grace pulled out a fine cable the width of a human hair and a transformer with an adaptor for the fine cable to plug into as the power supply.

The group watched as Grace ran a simple COBOL program. We didn't know it then, but we had just witnessed an early silicon chip-based computer. In Grace's opinion, the mainframe was

dead and would be replaced by 'multitudes of minicomputers' that would be linked by telephone lines, all working together. It was quite possible she had seen a demonstration of the U.S. Department of Defense's ARPANET, the precursor of the Internet.

In 1974, I moved to Australia. I was the only woman executive at AMP, the largest company in terms of assets. Part of my job was to buy all the hardware. Computers still used core memory, which was sensitive to heat fluctuations. Our Univac salesman told me there was a new type of memory called Metal Oxide on Silica (MOS), which was not affected by heat fluctuations. It cost \$1.5 million a megabyte. My boss was flummoxed, not only by the price but also by the size of the memory saying we already had two of the largest computers in the southern hemisphere, each with half a megabyte. I explained the system we were developing was to run real-time database systems, and more memory would be a distinct advantage. I was told if I could get the salesman to bring the cost down to \$1 million, we would buy it.

I wish I saved the invoice because it would be of great historical and hysterical value now. Today we refer to MOS memory as computer chips and a megabyte costs a fraction of a cent. Although manufacturing processes for chips have changed considerably, they are not very different from that first megabyte I bought for AMP in the mid 1970s.

And so it goes. Always new things to learn, always tremendous increases in technology accompanied by drop in costs. Always exciting new applications.

It's a wonderful career, especially for women. Come and join us.

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



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

DOI:10.1145/3484986

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

## Assessing Internet Software Engineering, Encouraging Competitions

*Andrei Sukhov considers the potential for reducing international tensions through competitive events, while Vivek S. Buzruk looks at the evolution of teaching Internet software engineering.*



**Andrei Sukhov**  
**Competitions,  
Not Confrontation**  
<https://bit.ly/3fFRNxV>  
May 21, 2021

In recent years, there has been an increasing tendency for confrontation between states and ideologies. The institution of sanctions, countersanctions, and the emergence of new hotspots occur with alarming regularity, reminiscent of the worst years of the Cold War. The level of tension is constantly increasing, and the meeting of preconditions for a decrease in the degree of confrontation is not generally within reaching distance. All this has had a negative impact on international scientific projects, including in the field of computer science.

The number of international competitions supported by government research funds has rapidly decreased.

New competitions involving collaborations between scientific teams of opposing 'sides' are no longer announced. These factors have led to a gulf between researchers from different countries, which does not contribute to the search for new knowledge or to the acquisition of communication and teamwork skills.

New collaborative research could be the bridge to allow the maintenance and development of communication between countries.

The question then arises, which international institutions are in a position to act as organizers of new competitions involving teams from different countries? Scientific foundations supported and funded by governments are often not well placed to take on this role, while international research communities united by professional interests could be seen as ideally suited to it. Organizations such as ACM and

IEEE include in their ranks researchers from a great variety of countries; these researchers successfully interact within the framework of these organizations. The influence of politics on the activities of such communities is still minimal; this allows ACM, for instance, to organize new competitions involving joint research by scientists in countries otherwise at odds.

From where can the necessary funding be sourced? ACM could act as founder of a fund specifically for this purpose. It would have to carry out the necessary preparatory work under its own auspices, but subsequently submit for wider discussion regulations regarding the activities of the fund. To improve transparency, it would be necessary to form a board encompassing all stakeholders on an equal basis. Such a board ideally would include ACM members representative of countries in conflict with each other, on a parity basis—as well as independent members. That is, on the one hand, representatives from the U.S. and other Anglo-Saxon countries; the European Union; and Japan should be included. And, on the other hand, representatives from China and Russia should also be at the table. In addition, largely non-aligned countries such as India, Brazil, Argentina, South Africa, Indonesia, and those of the Middle East should also be represented.

The board members also would have to represent universities, research institutions, and private companies. There should be restrictions regarding the representation of gov-

ernmental organizations—the fewer involved, the better.

It also would be desirable to reduce the direct contribution of governments to zero (or near zero) in the financial sphere as well; it might be necessary to refuse contributions to the fund by governmental organizations. Contributions from private companies would be preferred, but undue influence by any company would have to be avoided. To achieve this, it would be prudent to limit financial contributions made by any one organization to 10% of the total.

It is imperative that such a foundation be formed and start its activities as soon as possible, to prevent the further isolation of researchers living in different countries. I propose to conduct discussions to encourage comment and feedback, and when/if the idea is approved, to proceed with its practical implementation.



**Vivek S. Buzruk**  
**20 Years of 'Software Engineering for Innovative Internet Applications'**

<https://bit.ly/3ABd7k>

July 6, 2021

Twenty years ago at the Tenth International World Wide Web Conference, Hal Abelson and Philip Greenspun presented a paper on “learnings from teaching a Subject offered at MIT.”<sup>1</sup> Its subject was “Software Engineering of Innovative Internet Applications.”

I came across the paper in 2006/2007 while browsing the content of a similar subject/course. As a software professional, I was impressed by the focus and intentions of the paper, which emphasized “Engineering Software,” while teaching “Development of Innovative Internet applications” using current/emerging technologies.

The paper starts: “Why is software engineering part of the undergraduate computer science curriculum?” Industry expects the software developer to be also a great software engineer. How much does a student really practice software engineering during their undergraduate studies? Even if a student applies/uses software engineering principles during, say, a student project, are they evaluated based on their skills?

The paper suggests that in addition

to the core computer science curriculum, we also have to teach students:

- ▶ object-oriented design, in which each object is a Web service (distributed computing);
- ▶ concurrency and transactions;
- ▶ how to build a stateful user experience on top of stateless protocols; and
- ▶ about the relational database management system.

In 2021, these views need the attention of both academics and Industry. In the last 20 years,

- ▶ Internet Applications have evolved in size, complexity, and availability.
- ▶ Underlying front-end/back-end Web technologies and Internet Application execution platforms have undergone multiple changes.

▶ The evolution of technologies and platforms allowed software professionals to develop and support user-centric, multi-channel, secured, easily adaptable/integrable/accessible, intelligent, scalable, multi-tenant, Geography/Region-agnostic and real 24/7 high-performance Internet applications.

Yet, have they impacted the Vision and Mission of “Software Engineering for/of Innovative Internet Applications”? The title of the paper transcended changes in such technologies and platforms. I am sure content must have been modified every year. More importantly, in addition to technologies and platforms, students who attended this subject/course must have learned real-life situations/challenges, and the subject must have instilled the importance of Modularity, Distribution, Scalability, etc., through course content.

### Step Forward: Retrospective Action

When I recently got involved with an institute for its course curriculum of IT undergrads, I remembered Abelson and Greenspun’s paper. I recalled their smooth integration of various software engineering methods while teaching a student about building end-to-end (full-stack) Internet applications.

Today, during evaluation of IT undergraduate course curricula, I ask myself, considering that students learn “OO programming in XYZ,” “Web Technologies,” ... “Distributed & Cloud Computing,” etc., as independent courses/subjects:

- ▶ How can we ensure students will be able to apply a systematic, disciplined,

quantifiable approach to the development and maintenance of software using such paradigms and technologies?

- ▶ How they will be able to develop production-ready, maintainable software?

The question is more about time-window and habit. Industry will be overjoyed if it finds a Fresher who appreciates software engineering and, to some extent, the assimilated principles of software engineering.

Educational institutes try to match industry expectations and student aspirations within their own constraints. They cannot take responsibility for making every student a great software professional.

To achieve these goals within given constraints, few educational institutes think of logical threads running through their courses. One typical logical thread is continuously improving the student’s programming proficiency.

The missing dimension is the engineering approach of applying these isolated learnings to build applications; that is, students do not appear to learn “Software Engineering for XYZ Applications.” They are not evaluated on their software engineering approach during problem solving.

In fact, this small step of the software engineering dimension while educating students on various programming languages and new technology areas will make a big difference for students and industry.

In brief, Abelson and Greenspun’s paper needs attention from industry and Academics, who need to think about:

- ▶ What should be the Vision and Agenda in 2021?

▶ What, in addition to the core computer science curriculum, should academics be teaching IT undergraduates for “engineering innovative Information Systems of the next 20 years”?

### Reference

1. Teaching Software Engineering—lessons from MIT: by Hal Abelson and Philip Greenspun. Presented at the Tenth International World Wide Web Conference (Hong Kong), May 1–5, 2001: <https://bit.ly/3hce6gn>

**Andrei Sukhov** is a professor and head of the Network Security Research and Study Group of HSE University, Moscow, Russia. **Vivek S. Buzruk** is a DevOps consultant at SarvaTech Consultants Inc. India, and a ‘Board of Studies’ member for the IT Dept. of Pimpri Chinchwad College of Engineering (PCCoE) in Pune, India, mainly focusing on DevOps insights and custom software development.

© 2021 ACM 0001-0782/21/11 \$15.00



## ACM BOOKS

### Collection II

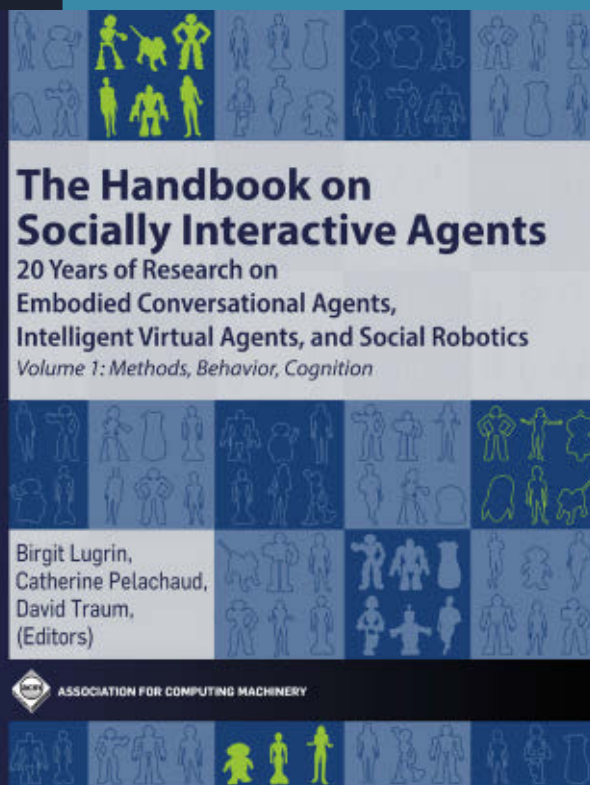
*The Handbook on Socially Interactive Agents* provides a comprehensive overview of the research fields of Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics. Socially Interactive Agents (SIAs), whether virtually or physically embodied, are autonomous agents that are able to perceive an environment including people or other agents, reason, decide how to interact, and express attitudes such as emotions, engagement, or empathy. They are capable of interacting with people and one another in a socially intelligent manner using multimodal communicative behaviors, with the goal to support humans in various domains.

Written by international experts in their respective fields, the book summarizes research in the many important research communities pertinent for SIAs, while discussing current challenges and future directions. The handbook provides easy access to modeling and studying SIAs for researchers and students, and aims at further bridging the gap between the research communities involved.

In two volumes, the book clearly structures the vast body of research. The first volume starts by introducing what is involved in SIAs research, in particular research methodologies and ethical implications of developing SIAs. It further examines research on appearance and behavior, focusing on multimodality. Finally, social cognition for SIAs is investigated using different theoretical models and phenomena such as theory of mind or pro-sociality. The second volume starts with perspectives on interaction, examined from different angles such as interaction in social space, group interaction, or long-term interaction. It also includes an extensive overview summarizing research and systems of human-agent platforms and of some of the major application areas of SIAs such as education, aging support, autism, and games.

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



## The Handbook on Socially Interactive Agents

*20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*

Edited by

**Birgit Lugin**

**Catherine Pelachaud**

**David Traum**

ISBN: 978-1-4503-8721-7

DOI: 10.1145/3477322

## Qubit Devices Inch Toward Reality

*Key questions and challenges remain, including how to scale qubit devices while reducing noise and errors to the point where the devices become useful.*

**T**HE MARCH TOWARD functional quantum computing devices has taken a long and winding road. Although the concept has been around since the late 1970s, when physicists Paul Benioff, Richard Feynman, and others began to explore quantum information theory, only recently have actual devices begun to take shape. Several companies, including IBM, have developed prototype quantum computing systems, while many research organizations have experimental devices in early stages of development.

Yet unlike classical computing, which has evolved over more than 70 years and is now mature, quantum computing, which harnesses quantum physics to leverage the uncertainty of a quantum state versus the certitude of a classical state, remains largely uncharted territory. An enormous amount of research is currently focused on ways to create or utilize quantum bits (“qubits”) to construct quantum mechanical systems that can harness physical events in nature to solve complex computing problems lying outside the practical grasp of classical systems. At the moment, qubit research remains in a



relatively nascent state and, as a result, it is not clear which approaches will ultimately prevail.

“There’s an enormous push to develop different quantum systems that could prove stable enough to be useful,” says Michael Cuthbert, director of the National Quantum Computing Center in Oxfordshire, U.K.

For now, qubit research is heavily focused on a handful of key technologies: superconducting circuits,

trapped ions, photonics, ultra-cold atoms, spins in silicon, color spins in diamonds, and an emerging area known as topological insulators.

Key questions and challenges remain, including how to scale devices while reducing noise and errors to the point where qubit devices become useful. Says Travis Humble, deputy director of the U.S. Department of Energy (DOE) Quantum Science Center at Oak Ridge National Laboratory, “We are

reaching a critical point where quantum computing technology is advancing rapidly. Over the next few years, we may see actual devices emerge that solve real-world challenges in drug development, financial modeling, cybersecurity, and physics.”

### A Quantum Challenge

Quantum computing devices are emerging in several shapes and forms—and they use radically different methods to process calculations. The common denominator is that all these systems use a quantum circuit to handle calculations. As with bits in classical computing models, qubits operate in a 1 or 0 state. Unlike classical computing, the information is challenging to record; moreover, the physics of the quantum realm unlocks the possibility of the occurrence of both 1s and 0s together. As a result, it is necessary to measure the qubit before and after a change, to understand its state at any precise moment.

Qubits typically maintain useful information for very brief periods of time, in some cases no more than a millisecond, as they are prone to high levels of noise and errors. By comparison, silicon can run a billion classical operations per second for a billion years before a statistical error occurs. “It’s possible to mitigate this instability through error correction,” Cuthbert says, “but the error correction comes with significant overhead. It’s necessary to reach somewhere around 1,000 physical qubits to get to the point where the error correction works effectively, creating a single logical or ‘forever’ qubit.”

So far, researchers and engineers have been able to build qubit devices to a scale of just over 60 qubits. What is needed are devices that can reach at least 1,000 qubits, or for a fully fault-tolerant error-corrected machine, 1 million qubits, experts say. Although better algorithms can address part of the noise and error problem by calculating more efficiently, they cannot address it entirely.

The goal, then, is to build qubit devices that break through today’s barriers and serve as components for quantum computers. While digital devices encode electrical signals in a string of ones and zeros, quantum systems re-

## Despite all the gains, the path to quantum computers remains lined with obstacles.

quire a highly stable two-level system that can establish a superposition of a one or zero, as well as the entanglement of states between different qubits.

This two-level system must bring quantum bits sufficiently close together to couple, in order for the device to generate usable output. In addition, a quantum computing device requires circuits that can detect and encode quantum states and generate data that can be used to understand events. The takeaway? “There are many different potential directions for quantum technology to take,” says Arun Persaud, a staff scientist at the Department of Energy’s Lawrence Berkeley National Laboratory.

### Qubit Devices Emerge

In 2019, scientists at Google announced they had passed a critical threshold in quantum computing: they managed to build a 53-qubit device that could solve in about 200 seconds a problem that would require upward of 10,000 years on the Summit supercomputer at Oak Ridge National Laboratory to complete. Although some in the computing community disputed the results, and it was widely acknowledged that the experiment did not solve any meaningful computing problem, many saw it as proof that quantum devices delivered what Google researchers then labeled “quantum supremacy.”

Google’s method relied on transmission line shunted plasma oscillation qubits, generally known as transmons, to calculate the problem. Transmons are a type of superconducting charge qubit that displays reduced sensitivity to charge noise. The same technique, which requires chilling superconducting metal to near absolute

zero, has been used by other commercial vendors to produce quantum computers. A transmon manipulates ratios between different energy levels inside of the superconducting device, and scaling up a system of these devices requires sophisticated engineering. “While today’s devices are relatively primitive in terms of the capacity and the number of instructions they can perform, there’s been remarkable progress with the technology over the last five years,” Humble says.

Another type of qubit device involves individual ions trapped in a stable configuration. These charged particles may consist of calcium, phosphorous, or other elemental atoms that are injected as a dilute gas into a vacuum system. An electrical or optical system, such as a laser, is used to induce coupling between the atomic quantum states. The technique has attracted the attention of several organizations, including Google and Honeywell, though these devices also remain in the tens of qubits, rather than the hundreds or thousands expected for commercial applications.

A third major area of research involves photonic systems. Instead of using electrons to store and carry information, these devices record information in the quantum state of light known as photons. The advantage of this approach is that researchers are familiar with the laser, which is itself a quantum technology. The challenge with the technique is getting photons to interact so the system can process the needed information. However, “It’s a very promising area because of the speed at which photons travel,” Humble says. “You can use them to distribute information between locations.”

Meanwhile, researchers continue to explore other types of qubit devices. For example, one early-stage technology uses topological insulators to encode information that is intrinsically resilient to the noise. The interior of these material systems functions as an insulator, while the surface contains conducting states that allow electrons to move easily. The exotic quantum states that arise are then used in a form of computation called ‘braiding’.

Like weaves in a fabric, braiding exchanges the relative ordering of the

electrons. However, due to their quantum nature, they perform meaningful calculations. Microsoft, which has partnered with the Quantum Science Center at Oak Ridge, has a leading commercial stake in the development of a topological quantum computer using ultra-narrow nanowire devices, and the laboratory is exploring new materials systems that could demonstrate these effects.

### Qubit by Qubit

The field of quantum computing took an important step forward in February, when Persaud and a team at Berkeley National Laboratory discovered a possible way to form self-aligned color centers that are close enough together to potentially push scalability to upward of 10,000 qubits—about two orders of magnitude greater than other ion-trapping technology.<sup>a</sup> The group accomplished this feat by using ion beams to generate artificial color centers in diamonds; it now is working on creating samples with single lines of color centers and measuring their quantum behavior. The color centers are microscopic defects that involve a nitrogen atom residing next to an empty space in the structure of the crystal.

By exciting passing ions to form nitrogen-vacancy centers in the diamond lattice, the centers and adjacent carbon atoms can serve as solid-state qubits. What's more, the crystal lattice helps protect their coherence and keep them entangled. This has significant ramifications because the quantum system can store data without requiring a cryogenic environment. "You can use optical detection of photons, a well-established technology, to detect and measure the state of qubits," Persaud says.

Also, in May of this year, a team at Washington University in St. Louis led by Jung-Tsung Shen, an associate professor in the Preston M. Green Department of Electrical & Systems Engineering of the university's McKelvey school of Engineering, developed a high-fidelity two-bit quantum logic gate based on light.<sup>b</sup> It bumps up efficiency by an order of magnitude using photonic dimmers, a type of photonic interaction

that exists in both space and frequency. When a single photon enters a gate, there is no reaction; however, when two enter simultaneously, it is possible to perform two-bit operations within a quantum framework.

Still another key breakthrough took place in June, when researchers at the Massachusetts Institute of Technology (MIT) found a way to reduce errors in two-qubit gates through fidelity advances in tunable couplers—systems that allow researchers to switch on and off various operations while preserving qubits.<sup>c</sup> In order to mitigate qubit-to-qubit interactions that lead to errors, researchers tapped higher energy levels within the coupler to cancel out problematic interactions, and ultimately, errors. The method "helps address one of the most critical quantum hardware issues today," says Youngkyu Sung, an MIT graduate student in electrical engineering and computer science.

### The Path Ahead

Despite all the gains, the path to quantum computers remains lined with obstacles.

Because decoherence takes place so rapidly, better material purity and lower noise in electronic systems are critical, says Erik DeBenedictis, co-chair of IEEE Quantum, an initiative that serves as IEEE's leading community for all the organization's quantum technology projects and activities. Many current systems simply produce too much heat, DeBenedictis says; as engineers attempt to scale quantum devices, "There are too many qubits pushed too close together and there's simply too much crosstalk."

Conquering the scaling problem will not be easy. Says Cuthbert, "Right now, the challenge is to produce an ensemble that's large enough and displays sufficient noise mitigation to provide useful results. We also need to develop better algorithms for use cases that demonstrate a clear quantum advantage."

Finally, there is a need to better understand different behaviors that take place on various quantum platforms, Cuthbert says. "Some are noisier than others. The task isn't only to discover

the perfect qubit and couple it with the next perfect qubit. There are issues involving connectivity between qubits and the speed at which those interactions take place." Adding to the challenge: different types of qubit devices and technologies have varying levels of tolerance to noise, something that may require different methods of error mitigation, correction, and control.

Nevertheless, the race to develop functional quantum computers continues. Over the next few years, qubit devices are likely to mature, scale, and evolve into far more advanced systems. For example, IBM has promised a 1,000-qubit machine by 2023.<sup>d</sup>

Says Persaud, "Quantum computing is a big field and there are many aspects to it. Right now, there's no consensus on which approach will win out. We may see one approach ultimately prevail, but we also might also see different types of systems for different quantum computing uses." ■

d <https://bit.ly/39eqRCv>

### Further Reading

Sung, Y., Ding, L., Braumüller, J., Vepsäläinen, A., Kannan, B., et al. **Realization of High-Fidelity CZ and ZZ-Free iSWAP Gates with a Tunable Coupler**, *Physics Review*, Vol. 11, issue 2. June 16, 2021; <https://journals.aps.org/prx/abstract/10.1103/PhysRevX.11.021058>

Lake, R.E., Persaud, A., Christian, C., Barnard, E.S., Chan, E.M., et al. **Direct Formation of Nitrogen-Vacancy Centers in Nitrogen Doped Diamond Along the Trajectories of Swift Heavy Ions**. *Applied Physics*, Volume 118, Issue 8. February 24, 2021; <https://aip.scitation.org/doi/10.1063/5.0036643>

Chen, Z., Zhou, Y., Shen, J., Ku, P., and Steel, D. **Two-photon Controlled-Phase Gates Enabled by Photonic Dimers**, *Physics Review A*, Volume 103, Issue 5. May 2021; <https://journals.aps.org/prx/abstract/10.1103/PhysRevA.103.052610>

Hays, M., Fatemi, V., Bourman, D., Diamond, C.S., Serniak, K., et al. **Coherent Manipulation of an Andreev Spin Qubit**, Cornell University, January 17, 2021. <https://arxiv.org/abs/2101.06701>

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

© 2021 ACM 0001-0782/21/11 \$15.00

a <https://bit.ly/3tQvWur>

b <https://bit.ly/3kzfOBU>

c <https://bit.ly/3zjy4fd>

# Holograms on the Horizon?

*Machine learning drives toward 3D imaging on the move.*

**R**ESearchers at the Massachusetts Institute of Technology (MIT) have used machine learning to reduce the processing power needed to render convincing holographic images, making it possible to generate them in near-real time on consumer-level computer hardware. Such a method could pave the way to portable virtual-reality systems that use holography instead of stereoscopic displays.

Stereo imagery can present the illusion of three-dimensionality, but users often complain of dizziness and fatigue after long periods of use because there is a mismatch between where the brain expects to focus and the flat focal plane of the two images. Switching to holographic image generation overcomes this problem; it uses interference in the patterns of many light beams to construct visible shapes in free space that present the brain with images it can more readily accept as three-dimensional (3D) objects.

"Holography in its extreme version produces a full optical reproduction of the image of the object. There should be no difference between the image of the object and the object itself," says Tim Wilkinson, a professor of electrical engineering at Jesus College of the U.K.'s University of Cambridge.

Conventional holograms based on photographic film can capture interference patterns that work over a relatively wide viewing range, but cannot support moving images. A real-time hologram uses a spatial-light modulator (SLM) to alter either the amplitude or phase of light, generally provided by one or more lasers, passing through it on a pixel-by-pixel basis. Today's SLMs are nowhere near large or detailed enough to create holographic images that can be viewed at a distance, but they are just good enough right now to create near-eye images in headsets



and have been built into demonstrators such as the HoloLens prototype developed by Andrew Maimone and colleagues at Microsoft Research.

A major obstacle to a HoloLens-type headset lies in the computational cost of generating a hologram. There are three algorithms used today to generate dynamic holograms, each of which has drawbacks. One separates the field of view into layers, which helps reduce computation time but lacks the ability to fine-tune depth. A scheme based on triangular meshes, like those used by games software that render 3D scenes onto a conventional two-dimensional (2D) display, helps cut processing time (although without modifications to handle textures, it lacks realism). The point-cloud method offers the best potential for realism, although at the expense of consuming more cycles. In its purest form, an algorithm traces the light emanating from each point to each pixel in the SLM's replay field. "Light from a single point can diverge to a very wide area. Every single point

source creates a sheet of refractions in the replay field," says Wilkinson.

A drawback of the point cloud is that light from every point will not reach every pixel in the target hologram, because it will be blocked by objects in front of it. That calls for software to remove the paths that should be occluded, which increases the number of branches in the code. Though it removes the need to map the light from every point onto every pixel in the SLM, the checks and branches slow down execution. Photorealistic holograms intended for use as codec test images, created using a method developed by David Blinder, a post-doctoral researcher at Belgium's Vrije Universiteit Brussel, and colleagues, take more than an hour to render using an nVidia Titan RTX graphics processing unit. However, numerous optimizations have been proposed that reduce arithmetic precision and the steps required, with some loss of quality, to achieve real-time performance on accelerated hardware.

The MIT approach uses several ap-



proximations and optimizations built around a deep neural network (DNN) made up of multiple convolutional layers that generate the image from many subholograms. This involves far fewer calculations than trying to map a complete point cloud directly to a final complete hologram. In conventional optimizations, lookup tables of diffraction patterns can help build those subholograms more quickly, but it is still an intensive process.

The DNN allows a more progressive approach to assembling the final image, which results in fewer calculations, particularly as the network can handle occlusion. The team trained the model on images of partially occluded objects and their sub-hologram patterns. The resulting algorithm can deliver images at a rate of just over 1Hz using the A13 Bionic accelerators in the iPhone 11 Pro. Without the computational optimizations provided by the DNN, the researchers suggest processing would take at least two orders of magnitude longer.

The MIT work underpins the need for good data in machine learning. The team looked at existing datasets for generating the required data, but all of them missed key components that made it impossible to train an effective model. One issue Ph.D. student Liang Shi and coworkers on the MIT project found is that existing datasets have objects clustered either at close range or far away from the viewer, with relatively few objects in the middle ground. This work needed a more consistent set of examples to avoid biases in the model that would lead to unwanted artefacts appearing in rendered scenes. Shi points out that the RGB image and depth data also need to be well aligned to ensure the DNN handles occlusions well. "This prohibits the use of real-world captured datasets, which often have undefined depth regions or misaligned depth values," he notes.

Wilkinson argues machine learning used in this way is unlikely to fit well with holographic displays that need to employ more extensive calculations of photon interference. These typically use Fourier transforms rather than an approximation of diffraction based on Fresnel optics, which underpin the subhologram algorithms.

"Machine learning is generally a

## The pixel density and resolution limitation of SLMs limit the effective viewing angle that can be supported, as well as the size of the "eyebow."

one-to-one or many-to-one translation process. Holography, because it's Fourier-based, is a one-to-many process. Each point can have an effect on every other," Wilkinson says. He points to that fact that the patterns seen in SLMs for full holograms tend to look like "random mush, though what you get out at the end is a lovely hologram. In these types of machine learning system, if you look at what they display on the SLM, you see a partially diffracted version of the real image."

Blinder says the approaches taken by MIT and others may not scale well if SLMs evolve to deliver larger fields of view. "The method is probably not suitable for holographic television with multiple viewers."

In the near term, this may not be an issue. The pixel density and resolution limitation of SLMs limit the effective viewing angle that can be supported, as well as the size of the "eyebow," which is the size of the region in which an observer will be able to see any of the hologram. Eye tracking coupled with rapid re-rendering can potentially compensate for these limitations in headsets and avoid the need to implement algorithms that can handle wider viewing ranges.

Machine learning also could help improve the perceived quality of the display's output. Gordon Wetzstein, an assistant professor of electrical engineering at Stanford University, says SLMs and the other optics in holographic displays are difficult to control, which leads to degraded image quality in experiments. "They al-

# ACM Member News

## COMBATING ONLINE MISINFORMATION



Filippo Menczer is a Distinguished Professor in the Luddy School of Informatics, Computing, and

Engineering at Indiana University in Bloomington, IN, where he also serves as director of the Observatory on Social Media.

Menczer, who received his undergraduate degree in physics from Italy's University of Rome, earned his master's degree in Computer Science and his Ph.D. in Computer Science and Cognitive Science at the University of California, San Diego.

After obtaining his doctorate, Menczer joined the University of Iowa as an assistant professor of management sciences. He later moved to the faculty of Indiana University, where he has remained.

Menczer's research is focused on analyzing and modeling the spread of information and misinformation in social networks, and on detecting and countering the manipulation of social media.

"I study all aspects of information diffusion, especially misinformation and manipulation of social media," Menczer said. This includes the detection of bots and the coordination of misinformation campaigns, as well as various kinds of abuse on social media by bad actors.

One tangent of his work is developing machine learning tools that will permit the public to detect and understand online manipulation.

Menczer thinks the challenges of online manipulation will not go away anytime soon. He hopes his research will help computing, public policy, and education increase the quality of information shared on social media, but without censorship or any other hindrance to free speech.

"I think that's going to be one of our top priorities: to create a healthier information ecosystem, and be less vulnerable to manipulation," Menczer said.

—John Delaney



## Advertise with ACM!

Reach the innovators and thought leaders working at the cutting edge of computing and information technology through ACM's magazines, websites and newsletters.



Request a media kit with specifications and pricing:

**Ilia Rodriguez**  
+1 212-626-0686  
acmm mediasales@acm.org



most never behave in exactly the way you simulate them. Machine learning can compensate for this difference by learning proxy models of the hardware," he says.

Wetzstein and coworkers used a camera-in-the-loop system to help train models to compensate for the optical imperfections and improve perceived image quality. Shi says the MIT team is working on similar approaches built on top of the DNN-based rendering system. "We have done follow-up work that takes into account both SLM deficiencies and a user's vision aberrations and compensate for both in the hologram computation."

Wilkinson reckons machine learning may be overkill in correction, at least for consumer displays. "Aberration is often quite a low-order problem, though there are some applications where it isn't, such as free-space optical communications. I would not be surprised if machine learning were ultimately used there."

The open question is whether machine learning will become a mainstay of holographic rendering, or whether work on algorithms will result in similar or even greater computational efficiencies that can be used in commercial holographic displays or projectors.

Wilkinson says opportunities remain for deterministic, rather than AI-based, techniques that are optimized for performance. In much of the computational holography work so far, he says there is a tendency to stick to known solutions for calculating holograms. "Today, there are just three algorithms we use for holography. That can't be right. We must be missing something here. We tend to find a solution that works and just use that. We don't think outside the box too much. I think that's a mistake."

One issue is that it is difficult to determine how well an algorithm performs in terms of image quality. This, says Wilkinson, is where machine learning's use of error minimization and norms may prove useful by providing automated ways of evaluating how close images are to a golden reference.

Blinder says holographic displays may take a similar path to systems such as nVidia's Deep Learning Super Sampling, which employs machine learn-

ing to interpolate higher-resolution imagery from low-resolution, partially rendered data. "Given the generality of DNNs, I think that hybrid systems will be the most likely future outcome. But this may be more challenging to achieve in holography since information is not well-localized spatially."

One possible direction for machine learning for holography may be in combining the output from multiple SLMs to try to build larger-scale projectors rather than headsets, Wilkinson says.

SLM size, resolution, and switching performance remain obstacles to delivering viable headsets, but work on computational holography has led to manufacturers taking more of an interest in these applications. "We are starting to see custom silicon appear that shows the manufacturers are taking holograms seriously," Wilkinson says.

With improvements in both hardware and algorithms, virtual reality may be able to move away from stereoscopic displays and the usability that go with them. □

### Further Reading

Maimone, A., Georgiou, A., and Kollin, J.S. Holographic near-eye displays for virtual and augmented reality, *ACM Transactions on Graphics*, Vol. 36, No. 4, Article 85 (2017). <https://doi.org/10.1145/3072959.3073624>

Shi, L., Li, B., Kim, C., Kellnhofer, P., and Matusik, W. Towards real-time photorealistic 3D holography with deep neural networks, *Nature*, Vol 591, p234, 11 March 2021. <https://doi.org/10.1038/s41586-020-03152-0>

Chang, C., Bang, K., Wetzstein, G., Lee, B., and Gao, L. Toward the next-generation VR/AR optics: a review of holographic near-eye displays from a human-centric perspective, *Optica*, Vol 7, Number 11 (2020). <https://doi.org/10.1364/OPTICA.406004>

Blinder, D., Ahar, A., Bettens, S., Birnbaum, T., Symeonidou, A., Ottevaere, H., Schretter, C., and Schelkens, P. Signal processing challenges for digital holographic video display systems, *Signal Processing: Image Communication* 70 (2019) 114-130. <https://doi.org/10.1016/j.image.2018.09.014>

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

# Filtering for Beauty

*Social media “influencers” use augmented reality filtering apps to appear more beautiful, together, and cool. Results may vary.*

**H**OUSTON-BASED HAIRSTYLIST Taylor Crowley, 36, has built a reputation as a social media influencer and has been using augmented reality (AR) filters for the past few years as a “confidence booster.”

“I’m a low-maintenance person. I don’t wear a ton of makeup because less is more,” she explains. “I try to choose filters that aren’t going to distort my face.”

Crowley is also “big into photography” and views image filters like a filter on a camera that can be used to change tonal qualities. “I keep it realistic.”

In one Instagram post of her posing with a large fish, Crowley used Adobe Lightroom to turn everything grayscale “because we live in Houston and fish in Galveston, and honestly, not everything is very pretty,” she says. “I thought grayscale made the fish pop out.”

Similarly, Crowley posted a picture of herself in a bathing suit on Cinco de Mayo and grayed out the background to accentuate the beer can she was drinking from—and her green bathing suit.

As someone who views content herself, “I think editing things that show a little more color or pop ... grabs my attention a little bit more.”

Crowley is quick to add that she does not use filters when she posts client-related content. “Because I’m a hair stylist, I feel that it’s cheating” to use filters, she says. “It’s not an honest reflection of my work. I also want people to have a reasonable expectation when they come to get their hair done.”

## The Instagram Influence

Ari Lightman, a digital media and marketing professor at Carnegie-Mellon University, says augmented reality filters have been around for several years and grew in popularity on the Instagram photo and video sharing social media app, which launched in 2010. They work by taking a video or image and transposing that on top of another video or im-



One of Facetune’s 10 tools for photo enhancement.

age, says Lightman, adding that he has not used any so-called “beauty” filters. However, as a “huge Star Wars fan,” Lightman says he has used an animated filter to put a Darth Vader helmet on his head when he wants to be funny.

Augmented reality (AR) is hot. ABI Research, a global tech market advisory firm, estimates the AR market in retail, commerce, and marketing will surpass US\$12 billion in 2025. Facebook, for one, is accelerating its efforts in the space; nearly one-fifth of the social media platform’s almost 10,000 employees are working on AR and Virtual Reality (VR) devices, according to a report in *The Information*.

Beauty filters are used by social media influencers trying to create a specific image. “They don’t want people to see how they look out of bed with no makeup,” Lightman says; they want to be perceived in a way that looks professional or creates a certain type of aesthetic.

Some filters allow people to “do things that are highly stylized” to their

appearance in videos or still images, Lightman says, like fixing bags under their eyes, reducing extra skin in the chin area, chiseling cheekbones, or reducing the size of their nose. “It gives the perception that this isn’t the real me, but a crafted view of how I want you to see me.”

Yet there can be a dark side to using filters. Teens, and particularly girls who use Instagram, blame that social media site for increased anxiety, depression, and suicidal thoughts, according to studies Facebook conducted, *The Wall Street Journal* reported in September. Facebook’s internal research reportedly found many of the problems are unique to Instagram because it focuses heavily on body and lifestyle.

Millennials tend to focus on “the perfectly curated images and they launched Instagram,” says Eric Dahan, CEO of Open Influence, a Los Angeles-based influencer marketing agency that claims to represent over 600,000 influencers and 1.2 million social media accounts.

Influencers generally use filters to enhance their appearance, he says. “We don’t ask them to do it,” Dahan says, but brands encourage it. If the agency is promoting a beauty brand, “we’ll work with beauty influencers and many are using augmented reality filters ... we see it all the time.”

Dahan says he is starting to see a “countermovement” to the use of beauty filters, in which people will post a picture using a filter, then post the same image without the filter to compare them and promote a body-positive message. “That’s a trend we’re seeing more of ... because of the pressure it puts on people” to put forth their best selves online.

Ellyn (a pseudonym), 22, was 14 when she downloaded her first beauty filter. While that is “still a very young and impressionable age,” she says, “there are 8-year-olds with Snapchat,” a phenomenon she finds disturbing. The social media messaging platform “catapulted the

whole era of filters and changing your appearance,” because users can apply filters to pictures that are meant to disappear after a few minutes, which was enticing, she says. “Now, there’s a million filters for anything—you can turn yourself into whatever you want.”

Some filters are designed to be used strictly for amusement. Snapchat allows people to put comical filters on top of their public personas, Lightman notes. “They are animated and cartoonish” and allow a user to create a digital avatar augmenting their actual image. Asked why people might turn to such animated filters, he says, “With everyone being on Zoom and the explosion of TikTok, people are hypersensitive about how they look.”

Lightman believes the massive increase in Zoom calls during the pandemic led more people to use filters, to hide a background or to reduce shadows (or too much light) on some faces.

Jeremy Bailenson, founding director of Stanford University’s Virtual Human Interaction Lab, says computer vision and deep learning networks can be combined to create “filters” that can dynamically change a video in real time. Bailenson thinks the most frequently used filter is the “Touch-up My Appearance” filter on Zoom, “which smooths out bags under eyes, covers up blemishes, and magically erases a bad night’s sleep,” he says. “Using features such as Animoji or Snap, one can take this to another level, for example, adding a mustache or a Unicorn horn.”

Lightman has no data on the use of filters, but believes they tend to be more popular with girls. Speaking from personal experience, he says both his daughters began using Instagram when they were approximately 14.

Regardless of gender, says Bailenson, “Once someone discovers the touch-up appearance filter on Zoom, it is hard to go back.”

### Popular Filters

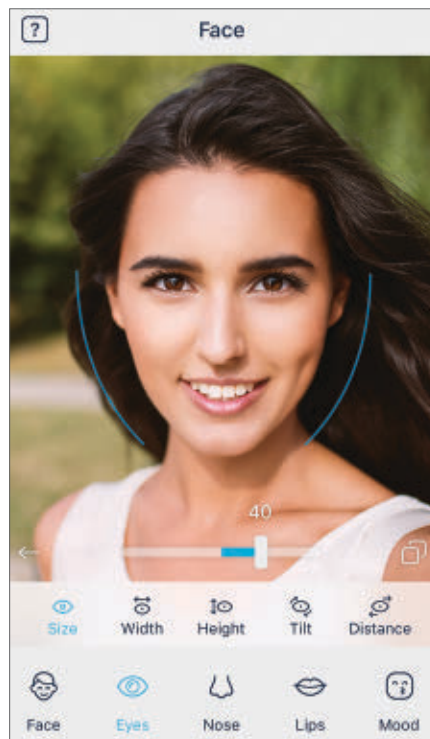
Ellyn most frequently uses the Facetune app to whiten her teeth or soften dark circles under her eyes; nothing drastic, she says. It also lets her fix lighting. “It’s not so much a filter, as opposed to an adjuster.”

She also uses VSCO, a photo-sharing app with preset filters that allows users

**“The escalation in procedures based on people looking at themselves on Zoom and trying to achieve some perfection, if you will, is a little problematic.”**

to enhance their images and videos, and Adobe Lightroom. “I use them to enhance what’s already there,” Ellyn says. “I don’t drastically change anything. If I’m looking at a photo and wish the backdrop looked little lighter, I can fix that and feel better about the photo.”

If she does not apply a filter, Ellyn says she sometimes feels “the picture is incomplete.” If a friend posts a picture she is in, Ellyn says regrets that they didn’t let her edit it first. “I’m so used to changing certain aspects of photos I don’t like, so when I see a picture of me completely unedited, I find myself wishing I could have,” she says. “But



Another Facetune beauty filter.

it’s a battle, because part of me says ‘that’s actually how I look and I should be happy with that and be confident.’”

Crowley likes an app called Retro Dust, “which gives everything a vintage feel,” she says.

### Pros and Cons

Beauty filters set an unrealistic expectation of what people look like, observers say. Call it the Kardashian effect: the family that built a billion-dollar empire promoting brands and often changing their appearances on social media.

“Their bodies and faces are so Photoshopped and cosmetically designed and a lot of them are seen as the beauty standard,” Ellyn says. “Young girls will look at that and say, ‘That’s what I should aspire to be.’ But that’s not a real version of that person, so I think people will aspire to be an edited version of themselves instead of embracing who they really are.”

Echoing Crowley, Ellyn says filters give people confidence. “I saw a video on Twitter the other day about a woman who had bunch of birthmarks, and filters and makeup allowed her to hide them and it makes her feel more comfortable,” she says. “In some ways, it does people a lot of good, but overall, it can be damaging to always see an edited version of someone.”

Some people use filters because they aspire to look like someone else and to get people to pay attention to what they post, Lightman says. The downside, of course, is that if they don’t look like the filters they’ve applied, “there’s a sense of authenticity that is diminished... We’re suffering from so much misinformation and so many deep fakes where people are literally putting other people’s images in situations they weren’t in or augmenting their voices.”

“There’s a fine line between photo editing and curated content, and being inauthentic with yourself and your followers,” Crowley says. “I have run into people I’ve followed for years and ... and they don’t look anything like what I thought. They’re not unattractive, but they have a whole different face.”

Plastic surgeon Simone Matousek in Sydney, Australia, says she often discusses the impact of filtered images in creating unrealistic expectations and driving

greater demand for plastic surgery. “Image filtering has led to people having a distorted sense of how they should look, what is achievable, and has largely driven the increase in cosmetic injectables,” she says. “A single picture on Instagram is very different to how that face may appear in reality.”

A “selfie” photo often distorts facial features, which is why most people do not look good in a close-range photo, Matousek says. “Even people considered some of the most attractive people in the world, such as models and celebrities, will rarely now upload an image without filtering, leading to unattainable beauty standards.”

Lightman says increasing numbers of women are coming out both for and against beauty filters. Cosmetics company websites will allow a user to apply a filter to see what they would look like without dark spots or blemishes, for example, which is helpful.

Like Matousek, Lightman says, “But when it gets into perverting self-image and unrealistic constraints and escalation of plastic surgery, I think that becomes a little unhealthy for society. Granted, there will always be people who see imperfections in themselves that want to fix them ... but the escalation in procedures based on people looking at themselves on Zoom and trying to achieve some perfection, if you will, is a little problematic. It’s an unrealistic expectation of beauty.”

### The Future of Filters

Trends come and go. Lightman thinks despite the pushback, beauty filters will always be used. However, with more people going back to offices and other places of business, there will be fewer videoconferencing calls, which will “right-size” the use of filters, he says.

“I think those beauty filters are not going to diminish completely, but they will become less sort of necessary,” Lightman says.

Bailenson sees that happening as well. “People engage in what psychologists call ‘self-presentation’ constantly in the real world, putting out the best version of themselves for a given context,” he says. “Clearly, self-presentation is amplified online. As the system evolves, we will find a balance for what type of filters are appropriate for work, play, dating, and other contexts.”

Dahan sees a transition happening, from a focus on beauty and cosmetics to one of self-care. Instead of striving for perfection, more frequently, the message is to be comfortable being yourself, he says.

“I think that’s been reflected within the influencer market,” and savvy brands are leading the way, Dahan says. “That’s what people are responding to. The beauty industry has been very toxic for so long and in some ways it still is, but if they want to continue to appeal to people, this is what they have to do.” **■**

### Further Reading

*ABI Research*

New Pandemic-Driven Sales Approaches Will Push Augmented Reality Revenues in Retail and Marketing to US\$12 billion in 2025, Feb. 3, 2021, <https://bit.ly/3yDCzS0>

*Eshiet, J.*

“Real me versus social media me.” Filters, Snapchat dysmorphia, and beauty perceptions among young women, 2020, California State University, San Bernardino, <https://bit.ly/3ARtgzR>

*Heath, A.*

The People With Power at Facebook as Its Hardware, Commerce Ambitions Expand, *The Information*, March 11, 2021, <https://bit.ly/3xsZi2V>

*London, L.*

How Beauty Filters Are Making Us ‘Look Better’ But Feel Worse, *Forbes*, March 23, 2020, <https://bit.ly/3wszoe4>

*Miller, M.*

Research looks at how Snapchat filters affect self-image, *Techxplore*, August 1, 2019, <https://bit.ly/2Vrmd0x>

*Oh, S., Bailenson, J., Kramer, N., and Li, B.* Let the Avatar Brighten Your Smile: Effects of Enhancing Facial Expressions in Virtual Environments, *Plos One*, Sept. 7, 2016, <https://bit.ly/3jYgedK>

*Rajanala, S., Maymone, M.B.C., and Vashia, N.* Selfies—Living in the Era of Filtered Photographs, *JAMA Facial Plastic Surgery*, Nov. 15, 2018, <https://bit.ly/3AIyIET>

*Robin, M.*

How Selfie Filters Warp Your Beauty Standards, *Teen Vogue*, May 25, 2018, <https://bit.ly/3hnHVLk>

*Sahoo, S.*

Beauty Trap: A study on impact of beauty filters on millennial women on Instagram. 2019. Christ University, Bangalore, <https://bit.ly/2T2A7oY>

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

© 2021 ACM 0001-0782/21/11 \$15.00

### ACM News

# ACM Committee Proposes RTA System Guidelines

The Association for Computing Machinery’s U.S. Technology Policy Committee (ACM USTPC) has released a Statement on Principles for the Development and Deployment of Equitable, Private, and Secure Test Administration Systems, proposing guiding policies and principles for the design, use, and oversight of Remote Testing Administration (RTA) software of the kind widely used to proctor the exams of millions of students globally.

The Statement outlines several guiding principles for those who develop and provide remote test administration software, recommending that developers and providers of RTA systems strive to:

- ▶ ensure equitable outcomes for marginalized learners;
- ▶ use end-to-end encryption;
- ▶ guarantee data collection is targeted, minimized, and transparent;
- ▶ not access local data on a test-taker’s computer;
- ▶ voluntarily share all pertinent information when determining that someone was involved in academic misconduct;
- ▶ assure their systems are accessible to all potential users, including users with disabilities, and those who have limited equipment or weak Internet connectivity; and
- ▶ develop uniform benchmarks and certification procedures to assess and document the comparative effectiveness of RTA systems in identifying students receiving unauthorized help.

“Automated test monitoring technology that observes a student’s behavior has become widespread,” says ACM USTPC chair Jeremy Epstein, adding that such technology “is opaque, and may introduce additional bias as well as privacy risks. The ACM USTPC is ahead of the curve in putting forth principles that can be applied to test-taking software. Our goal is to help school systems and universities know what questions to ask in acquiring and using such systems, and to help providers of such systems think about characteristics to include in the design of their products.”



DOI:10.1145/3486628

Pamela Samuelson

# Legally Speaking

## Text and Data Mining of In-Copyright Works: Is It Legal?

*How copyright law might be an impediment to text and data mining research.*

**T**EXT AND DATA MINING (TDM) uses statistical analysis tools to extract new knowledge from large quantities of text or data for purposes by finding patterns, discovering relationships, and analyzing semantics. It is used in a wide variety of fields from biomedical research to digital humanities. Copyright poses no obstacle to TDM research as long as the corpus of text and data being analyzed consists solely of public domain works.<sup>a</sup> Copyright may, however, be a barrier to TDM research as to vast arrays of in-copyright works created in the past century.

This is because copyright regulates making copies of protected works and TDM requires researchers to make several types of copies during different stages

of the process: from scanning copies of analog works to formatting the texts and data to preparing them for processing to extract useful information from the vast quantities being searched to storing the data after mining is completed.

Governments that aspire for their industries to become global leaders in artificial intelligence (AI) fields are beginning to realize their knowledge econ-

**Under the amended law, users are allowed to analyze in-copyright works for machine learning purposes.**

omies are more likely to thrive if they allow researchers to make copies of in-copyright works for TDM purposes. U.S. appellate courts have enabled this by ruling that TDM copying of in-copyright works is not infringement. Japan has enacted laws to allow TDM research copying. The E.U.'s 2019 Directive on Copyright and Related Rights in the Digital Single Market (CDSM) has mandated member states must adopt copyright exceptions for TDM research purposes.

### U.S. Fair Use TDM Decisions

Two U.S. appellate court decisions—*Authors Guild v. Google* and *Authors Guild v. HathiTrust*—have ruled that copying of in-copyright texts for TDM research purposes was fair use, not infringement. These lawsuits grew out of the Google Book Search Project (GBS).

GBS is a corpus of millions of digital books to improve its search technologies that Google developed after making a deal with the University of Michigan in 2004 to scan all eight mil-

<sup>a</sup> In the U.S., any work published before 1926 is reliably in the public domain. In other countries, copyright terms that last for the life of the author plus 50 or 70 years make it more difficult to determine whether works are in the public domain.



lion books in its library's collections. In return, Michigan got back from Google digital copies of the books it scanned. Google struck similar deals with several other state-related universities. The HathiTrust digital library was formed to host a collection of library digital copies Google provided to Google's state-related library partners.

By 2005, Google had digitally scanned millions of books from research library collections, the overwhelming majority of which were in-copyright. Later that year, the Authors Guild and three of its members brought a class action lawsuit charging Google with copyright infringement for making these digital copies.

From the Guild's perspective, Google's systematic copying of the entire contents of millions of all types of in-copyright books for commercial purposes was completely unjustifiable. The main norm underlying copyright ownership is that people who want to make copies of authorial works must ask for and get permission to make such copies, which Google did not do.

Google defended by saying its copying of the books was fair use because its purpose in scanning the books was socially beneficial. It was necessary to copy the

entire contents to index the books' contents, serve up snippets in response to user search queries, and enable Google to engage in non-consumptive research (for example, creating the Ngram viewer to enable users to see trends in word and phrase usages over time and improving its translation tools).

Google also asserted the snippets it served up were fair use because they were too few in number and too short in length to have harmful impacts on markets for the books. People do not use GBS to consume book contents. GBS searchers are generally looking for facts books may contain (for example, 'How many buffalos are there in Yellowstone National Park?') and copyright does not protect facts. Indeed, because Google provided links to sites at which users could purchase books responsive to user search queries, it was more likely GBS would benefit the market for books, not harm it.

An appellate court found Google's arguments more persuasive than the Authors Guild's claims. It observed GBS had enabled new kinds of research to be undertaken, specifically mentioning TDM as an example. Research and scholarship are two of the statutorily

avored fair uses, so this too supported Google's defense.

The *HathiTrust* decision more directly addressed TDM research issues. HathiTrust allows researchers from consortium member institutions to conduct searches across its corpus of millions of books (now totaling approximately 17 million volumes) to identify every book mentioning the person, place, or phenomenon for which researchers were looking.

HathiTrust provides researchers from partner institutions with bibliographic information about specific books in which the referent search term appeared and even data about page numbers where the referents could be found. The court considered this beneficial research purpose to strongly favor HathiTrust's fair use defense.

### Japan's Special TDM Exception

Recognizing how important TDM is to achieving success in AI fields, the Japanese legislature adopted a special exception to copyright rules to enable TDM research in 2009. It was the first nation in the world to enact such a law. Yet, AI researchers complained this exception did not fully address the needs of TDM

and AI researchers, so in 2018 Japan amended its copyright law to respond to those concerns.

Under the amended law, users are allowed to analyze in-copyright works for machine learning purposes. As long as TDM researchers do not exploit the protected expression in the works, but only process the data to extract knowledge, they do no harm to the legitimate interests of copyright owners whose rights extend only to control exploitations of expressive aspects of their works. It is thus fair game to feed in-copyright works as raw data into computers to process it for deep learning purposes.

The amended law also permits researchers to make incidental digital copies of works for TDM purposes. This recognizes that incidental copies are necessary to carry out machine learning activities. This too causes no harm to copyright owners' legitimate interests.

An additional provision of the amended law allows TDM researchers to use digital copies of in-copyright works for data verification purposes. The legislature recognized this kind of use is important to enable researchers to ensure their results and insights from TDM research are sound. This activity too is not detrimental to the legitimate interests of copyright owners.

### TDM Exceptions in the CDSM

An early draft of the European Commission's proposed CDSM directive would have required member states of the E.U. to adopt a new copyright exception to allow researchers at nonprofit scientific organizations to engage in TDM research as long as they had lawful access to the databases on which they conduct their work. This new exception was to be mandatory as well as non-waivable by contract.

The final Directive, which E.U. member states were supposed to have implemented the TDM exceptions in national laws by June 2021—although not all have done so—authorizes the TDM research exception to apply to nonprofit cultural heritage researchers as well as to scientific researchers.

In response to concerns that limiting the TDM exception to nonprofit researchers would undermine E.U.'s aspirations for their industries to build AI systems that could compete in the global marketplace, the Commission was

## Downloading Sci-Hub would be a risky strategy for TDM researchers who do not want to be sued for copyright infringement.

persuaded to add a second mandatory TDM exception for other researchers, including those engaged in commercial TDM research. However, this exception can be overridden by contract by owners of databases on which these researchers want to engage in TDM analysis.

Some scholars have expressed concerns the CDSM TDM exceptions, while steps in the right direction, will prove to be too narrow and uncertain in scope to fully address the needs of TDM researchers. Japan's more capacious TDM-enabling rules would be more responsive to researchers' needs.

### TDM on Sci-Hub's Corpus?

Sci-Hub is a well-known repository of vast quantities of the world's scientific journal literature, much of which is usually kept behind proprietary paywalls. Publishers such as Elsevier have sued Sci-Hub and its founder for copyright infringement. Courts have held this database contains much infringing materials and has forced its founder to shut it down. However, Sci-Hub's corpus has reemerged as a resource for scientists and can still be easily found on the Internet.

Many researchers would like to use it for TDM purposes, but is this legal?

The desire to use Sci-Hub for TDM research arises in part because numerous proprietary publishers of scientific journals offer institutional database subscriptions to universities and other research institutions that are not cross-platform interoperable. Researchers consequently cannot run searches across various proprietary databases. Cross-publisher collaborations are rare.

Moreover, the license terms on

which proprietary databases are available may impair researchers' ability to make the full use of TDM tools. Publishers and some collecting societies are promoting licensing of TDM as a value-added service for which research institutions should pay. Some licenses are more restrictive than TDM researchers would want.

Even scientific researchers who work at institutions that subscribe to proprietary databases want to use Sci-Hub to do TDM research. That database is easier to use than some of the publisher repositories. The Sci-Hub database is far more comprehensive than any of the proprietary databases. And there are no license restrictions to limit researcher freedom to investigate with TDM tools to their hearts' content.

Downloading Sci-Hub would be a risky strategy for TDM researchers who do not want to be sued for copyright infringement. But running TDM searches on the Sci-Hub collection hosted elsewhere involves only the kind of transient copying that the U.S. courts have found too evanescent to be an infringing "copy" of copyrighted television programming. The results extracted in the course of doing TDM research on Sci-Hub would be unprotectable facts.<sup>1</sup>

Consequently, it is conceivable TDM researchers would not infringe U.S. copyright law if they used Sci-Hub for TDM research purposes. However, the E.U. exceptions allowing TDM research are predicated on researchers having lawful access to the text and data they mine.

### Conclusion

Only a few countries in the world have flexible fair use or fair use-like exceptions to copyright rules that would enable them to use this tool to justify TDM research copying. Hence, legislation will be necessary for allowing TDM researchers to take full advantage of this new suite of tools to expand the horizons of what can be known from digital explorations of large corpora of data and text. **□**

### Reference

1. Carroll, M. *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful* 53, UC Davis L. Rev. 893 (2020).

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

Copyright held by author.



## Privacy

# Privacy Engineering Superheroes

*Privacy engineers are essential to both preventing and responding to organizational privacy problems.*

**D**OES YOUR ORGANIZATION want to offer cookie choices without annoying popups? Do you want to share sensitive data in aggregate form without risking a privacy breach? Do you want to monitor data flows to ensure personal information does not end up in unexpected places? What if personal information does leak out and now you need to clean up the mess? Do you want to do this in the messy, failure-prone world of a large system? Who ya gonna call? How about a privacy engineer!

The privacy profession is dominated by lawyers—who certainly play a critical role—but privacy engineers are often the real superheroes when things go wrong, and essential to preventing privacy disasters. Privacy engineering has emerged as a growing discipline focused on finding practical and often technical solutions to privacy protection. Organizations hire privacy engineers to develop privacy-protective products and services, build tools to promote and monitor privacy compliance throughout their organization, and to detect and remediate privacy problems. Privacy engineers may play a holistic role or focus on specific areas such as front-end, back-end, user experience, product management, or legal compliance.<sup>1</sup>

One of us (Lea Kissner) is a privacy engineering practitioner who has led privacy engineering teams at four companies, and one of us (Lorrie Cranor) has spent almost 20 years in



academia and now co-directs an academic program to train privacy engineers.<sup>a</sup> Over dinner a few years ago, Kissner complained about the lack of a venue for privacy engineering practitioners to discuss their problems and solutions and learn about research that could be applied to their work. Cranor noted that privacy researchers would also benefit from learning more about actual problems faced by practitioners and from having a forum where they could share their research results directly with practitioners. By the time we finished din-

ner, we had a skeletal plan for a new conference.

### Privacy Engineering Practice and Respect

We started the conference on Privacy Engineering Practice and Respect (PEPR) in 2019,<sup>b</sup> bringing together privacy engineers from academia, industry, civil society, and government to share their expertise. Privacy engineers from industry have discussed ways data deletion can fail in truly bizarre ways in large-scale systems, while academic researchers have presented user study

a See <http://privacy.cs.cmu.edu>

b See <http://pepr.tech>

results providing insights into why users do not seem to understand many privacy-related icons. PEPR brings us all together to discuss privacy-related ideas and how they work (and fail) in practice. Because the traditional academic paper format is focused on conveying research results rather than experiences from practice, and because many of the practitioners we want to hear from are not experts at writing academic papers, PEPR asks prospective speakers to submit talk outlines rather than papers. Since 2020, all PEPR talks have been recorded and made freely available after the conference.<sup>c</sup>

PEPR 2021 was fully remote, but was probably the largest (virtual) gathering of privacy engineers ever, with over 500 participants attending talks and engaging in discussions. PEPR focuses on building respectful products and systems instead of breaking them. It is easy to look around and see the ways things are broken. It is easy to succumb to nihilism. But we want to build the world we want to live in and sys-

<sup>c</sup> Slides and recordings available at <https://bit.ly/3tx6j1l> and <https://bit.ly/3hkUW7R>

## Privacy engineering has a fundamentally different focus than much of the privacy field as a whole.

tems will not get better unless we build them better. So while we are interested in both breaking and building, we lean toward building.

For the same reason, we look more broadly than privacy and also focus on respect. According to Wikipedia, “Respect is a positive feeling or action shown toward someone or something considered important, or held in high esteem or regard; ... it is also the process of honoring someone by exhibiting care, concern, or consideration for their needs or feelings.”

Building respectful systems requires

considering concepts beyond privacy, including security, trust, safety, algorithmic fairness, and more. We must move out of our disciplinary silos and consider our work in a broader context.

Consider for example the issue of filtering out unwanted email. If one looks at it from the perspective of privacy, the first questions tend to be around whether it should be done by default and whether it is reasonable to build models over the contents of multiple users’ inboxes. But it is also a security issue: some of those unwanted email messages are direct security threats such as phishing or messages that contain malware. And without scanning large amounts of email to understand the quickly changing threat landscape, it is effectively impossible to identify those threats. However, building models based on everyone’s personal email may be problematic. Are these models fair or do they operate unfairly toward people using, say, particular dialects?

Privacy engineering has a fundamentally different focus than much of the privacy field as a whole. Many organizations are interested in privacy because they want to be compliant with

# ICCQ

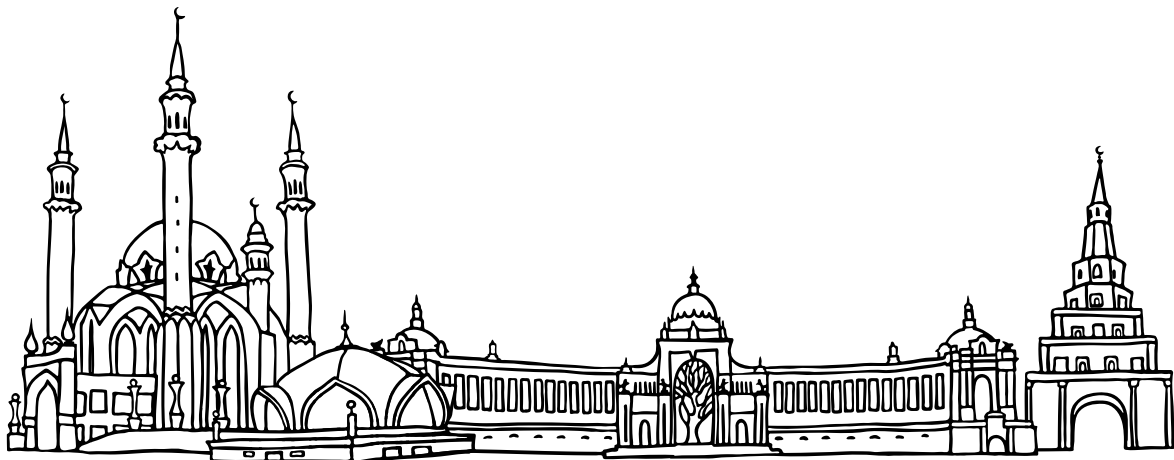
The Second International Conference  
on Code Quality (23 Apr, online)

Static/Dynamic Analysis, Program Verification,  
Bug Detection, and Software Maintenance

[www.iccq.ru](http://www.iccq.ru)

CfP closes on 18 Dec

In cooperation with  
ACM SIGPLAN and SIGSOFT  
IEEE Computer Society



laws such as GDPR, CCPA, and a whole alphabet soup of regulations coming into effect. While we certainly want systems to be built and operated in accordance with the applicable laws, that should be a side effect of building privacy-respectful products and systems in the first place. Compliance is necessarily reactive. It is responsive to failures of the past. If you are doing new things, then you are likely to hit new failure modes—compliance is not going to be sufficient. For one, when things go really wrong, no one cares about paperwork. But also those laws are moving; if you have built your systems around the checklist from compliance, then it is unlikely to be able to smoothly incorporate changes to the rules along with changes to the threat models your users face. Proactive privacy engineering considers how the changing world impacts your strategic direction to help shape your products and systems to better support your users and the folks affected by your systems.

### Privacy Engineering Specialties

Addressing privacy engineering requires many skillsets—here are some of the major specialties and roles that have developed in this young field.

**Analysis/consulting.** These folks look at your product/system (or better yet, the plans you have to build one), ask questions, find failures before they happen, and help you design in a way to robustly avoid those failures. For example, someone develops a new feature and the analyst asks questions such as “How can the user delete this information and is it really deleted?” and “If we put this nifty crypto in there, we can avoid collecting this information at all. Does that open up abuse vectors?” Privacy analysis/consulting folks have a skillset somewhat akin to security reviewers, but usually with a heavier emphasis on how humans of various stripes interact with each other and your product. They may well audit code.

**Privacy products.** This is where you are building the privacy technology users can see, such as account deletion pages, interfaces that let users see and control their data, and so forth. It is usually best when privacy affordances are part of the main product rather than a standalone tool.

## While we are interested in both breaking and building, we lean toward building.

**Math and theory.** Need to do anonymization? Need to analyze whether there is some kind of funky joinability risk across all your datasets? Need to figure out what is going to happen when you delete a particular type of data? Is it going to break your abuse models? Math.

**Infrastructure.** If you have got infrastructure, you probably need privacy infrastructure. For instance, when you want data to be deleted, you need a system to kick that off and then monitor it. You need access control and probably something to deal with cookies. Privacy engineers who build infrastructure have the infrastructure-building software engineering skills as well as knowledge of privacy.

**Tooling and dashboards.** You might have a data deletion or access system, but how do you help the humans in your organization understand what is going on in there or get access? Tooling! Tooling and dashboards are also extremely useful for things like efficient, accurate analysis and review of systems. Good remediation tooling holds a mirror up to the rest of the organization and tells them both how they are doing, exactly what they can do to get better, and how much better they are expected to be.

**User experience (UX).** The difference between a privacy-respectful product and a privacy-invasive product can sometimes be a matter of user experience. Is there transparency about what information is being collected and how it will be used and are users able to easily understand and implement privacy choices? A lot of privacy-engineering UX focuses on design and testing of privacy-related affordances (settings, dashboards, dialogues, icons, and so forth). There is also a need for privacy engineer involvement

in writing clear and accurate privacy notices—ideally this is not a job left only to the lawyers! Privacy UX research uses qualitative and quantitative methods to study people and how they interact with a product.

**Privacy policy.** This is not always done by privacy engineers, but we strongly recommend having someone with privacy engineering skills in the mix. It is very easy to write aspirational policy or policy that does not work with systems as they are. Both of these are bad. Kissner has written and managed privacy policy at both large scale (Google) and small scale (Humu).

**Privacy process.** Process design is key to getting things done. A deeply engineering-integrated process that aligns the incentives of everyone involved is key to making privacy a “well-lit path,” something smooth and efficient for the rest of the organization. If you make the rest of the organization grumpy, you are not going to get good results.

**Incident and vulnerability response.** Things will go wrong in a real system. There will be bugs and process failures and new issues you have never even thought of before. Incident responders use a cool head, excellent communication skills, and knowledge of how things go wrong in privacy to help folks work out what has gone wrong, how to fix it, and then how to keep that whole class of failures from happening again.

Of course, privacy engineers do not work in a vacuum, and they must collaborate with lawyers, policy makers, software developers, and many others. With all these roles and specializations, there are many opportunities for all kinds of privacy engineers with diverse skillsets to help organizations build privacy into their products and services and help save the day when things go wrong. ■

### Reference

1. Brook, B. The disciplines of modern data privacy engineering. IAPP. (Sept. 9, 2020); <https://bit.ly/3DZv774>

**Lea Kissner** ([lkissner@twitter.com](https://twitter.com/lkissner)) is Head of Privacy Engineering, Twitter, San Francisco, CA, USA.

**Lorrie Faith Cranor** ([lorrie@cmu.edu](mailto:lorrie@cmu.edu)) is Director and Bosch Distinguished Professor in Security and Privacy Technologies, CyLab Security and Privacy Institute and FORE Systems Professor, Computer Science and Engineering & Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.

Copyright held by authors.

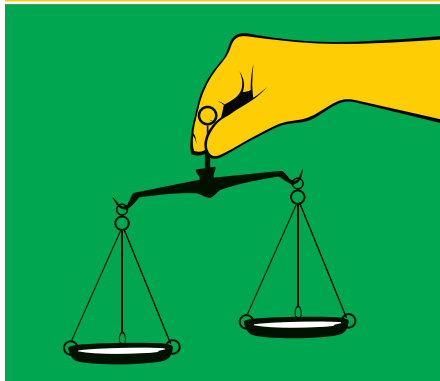
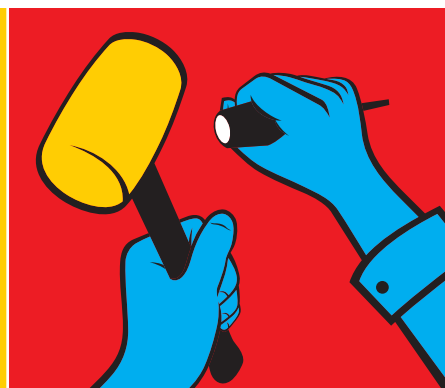
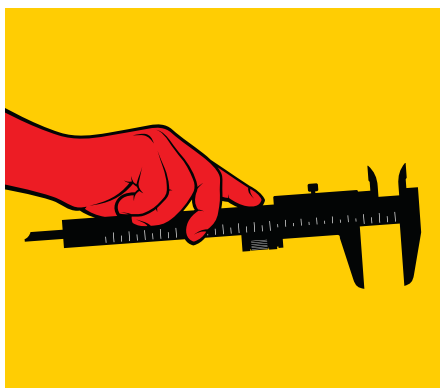
► Susan J. Winter, Column Editor

## Computing Ethics Shaping Ethical Computing Cultures

*Lessons from the recent past.*

**P**UBLIC CONCERN ABOUT COMPUTER ethics and worry about the social impacts of computing has fomented the “techlash.” Newspaper headlines describe company data scandals and breaches; the ways that communication platforms promote social division and radicalization; government surveillance using systems developed by private industry; machine learning algorithms that reify entrenched racism, sexism, cis-normativity, ableism, and homophobia; and mounting concerns about the environmental impact of computing resources. How can we *change* the field of computing so that ethics is as central a concern as growth, efficiency, and innovation? There is no one intervention to change an entire field: instead, broad change will take a combination of guidelines, governance, and advocacy. None is easy and each raises complex questions, but each approach represents a tool for building an ethical culture of computing.

To envision a culture of computing with ethics as a central concern, we start with the recent past, and a subdiscipline—computer security research—that has grappled with ethics concerns for decades. The 2012 *Menlo Report*<sup>3</sup> established guidelines for responsible research in network and computer security. After *Menlo*, new requirements for ethics statements in computer security and network



measurement conferences illustrate the use of governance for centering ethics in computing. Historically, a volunteer organization, Computer Professionals for Social Responsibility (CPSR), engaged in advocacy beginning in the 1980s to shape a more ethical future of computing, and influenced many of today’s leading Internet watchdog and activist groups.<sup>4</sup>

Each of these efforts represents a different way of doing ethics beyond the scale of individual decision mak-

ing, and each can be adapted in areas such as data science, social media, IoT, and AI. But as shown in Table 1, these cases also illustrate the difficult questions, trade-offs, and compromises required for culture change, and the challenges of work left to be done. Moving beyond the reactive “techlash,” tech workers and computing researchers interested in systemic ethical change can be inspired by these efforts while appreciating the trade-offs and understanding the

uphill nature of organized, sustained, and collective ethics work at scale. To illustrate each method, three examples are described in Table 1.

### Setting Research Guidelines: *The Menlo Report*

Recent calls for codes of ethics for data science and social media research echo similar concerns that roiled computer security research in the 2000s. In response, the U.S. Department of Homeland Security (DHS) organized funded researchers and invited legal experts to collaboratively develop guidelines for ethical network security research. Our interviews with 12 of the 15 primary *Menlo Report* authors found the effort made smart use of existing resources, including funding for a related research program and existing ethical guidelines from other domains. But the authors faced at least two difficult challenges. First, *who* should set ethical guidelines for a field? Because the *Menlo* work was involved and long-term, it largely fell to a group already funded under a DHS network security program. Second, *how* does a volunteer group set guidelines that people know about, ascribe to, and follow? The *Menlo Report* did not produce large-scale regulatory changes and authors we spoke with lamented the lack of resources for long-term education and training to support and evaluate the Report's impact.

The limited reach of *Menlo* is demonstrated in persistent computer security research controversies. Recently, cybersecurity researchers at the University of Minnesota caused an uproar with research that exposed vulnerabilities in the socio-technical system for approving Linux patches. Though their aim was to study Linux contributors' ability to detect security vulnerabilities, they believed their research did not involve human subjects (a judgment with which their Institutional Review Board agreed). The Linux community, however, reacted with anger reminiscent of the fallout from the famous Sokal Hoax, calling the work a "bad faith" violation of the community's trust.<sup>8</sup> This case illustrated exactly the kinds of uncertainty at the intersection of humans and systems that *Menlo* was written

## Recent calls for codes of ethics for data science and social media research echo similar concerns that roiled computer security research in the 2000s.

to address. Following *Menlo* guidance might have helped the researchers craft a clearer statement of their ethical deliberations and decisions, and might have helped the IRB identify the human stakeholders at the center of the research. As this example illustrates, expanding the reach of guidelines like the *Menlo Report* is a formidable challenge.

### Research Governance: Conference Ethics Statements




Another model computer security researchers are using to create more effective organized, sustained, and

collective action is to build ethical guidelines into gatekeeping processes. Conference peer review can help govern research ethics by only publishing work that meets a higher standard, effectively defining ethical reflection as a necessary part of security research processes. To encourage researcher reflection and compliance, many of the top computer security and network measurement conferences now require an explicit declaration of ethical considerations. In 2012, USENIX, one of the top conferences in computer security, included a requirement in their Call for Papers that researchers "disclose whether an ethics review ... was conducted and discuss steps taken to ensure that participants were treated ethically." Other important security and network measurement conferences soon followed.<sup>a</sup> After instituting these requirements, more conference papers now discuss ethical research practices.

But post hoc reflection and conference reviewing alone does not ensure ethical research—this governance takes place after the work is done. In a recent survey of computer security re-

a Network and Distributed System Security Symposium (NDSS), IEEE Symposium on Security and Privacy (Oakland), the ACM Conference on Computer and Communications Security (CCS), and the Conference for the Special Interest Group on Security, Audit and Control.

**Table 1. The work and challenges of ethical change methods at scale.**

	 <b>Guidelines:</b> Menlo Report	 <b>Governance:</b> Security conference ethics statements	 <b>Advocacy:</b> CPSR
<b>Ethics Work</b>	▶ Adopting principles and norms	▶ Reflection, rethinking and changing research design, writing	▶ Co-education, writing, publishing in popular press
<b>Hard Questions</b>	▶ Who sets the principles and norms (and what expertise is needed?) ▶ How to make people notice and follow guidelines without enforcement mechanisms?	▶ Who educates researchers? ▶ Do discussions of ethics in scholarly papers enhance ethical research? ▶ How to incorporate voices of stakeholders underrepresented among researchers?	▶ How to get people involved despite professional risks? ▶ How to keep volunteer efforts focused on achieving similar goals? ▶ How can computer professionals act responsibly?

searchers we conducted, a majority of respondents remain concerned about ethical practices in their community. Many computer security researchers engage legal experts, but lawyers are not well-positioned to help researchers work through ethical conundrums and what is legal is not always ethical. Instead, we found most computer security researchers learn about ethical research through interpersonal sources, such as graduate supervisors and university colleagues, so frequent and respectful discussions of ethics among colleagues are important.<sup>2</sup> In many computer security research labs these discussions are ongoing, but more need to take place.

Narrow perspectives are another important issue facing research governance. Though members of marginalized communities are frequently unfairly impacted by technical systems, they are too often underrepresented on program committees and guideline-setting bodies. Consequently, narrow perspectives restrict the frameworks, methods, and remediations that researchers consider in both developing systems and in designing governance instruments like conference policies. In addition to broadening participation in computing, classroom education can help introduce disparate technological impacts and train future computer researchers to “attune”<sup>1</sup> their work to issues of power, exclusion, and inclusion. Collectively, tech workers and computing researchers can change the culture of computer science by developing policies that both empower and are informed by people who are marginalized in technology research and development.

**Advocating for Limits: Computer Professionals for Social Responsibility**

Building a sustainable advocacy organization is a third model for collective action. Computer Professionals for Social Responsibility (CPSR) was an early exemplar. CPSR started in 1981 as a listserv at Xerox PARC, incorporated as a non-profit in 1983, and soon grew in size and influence, with chapters across the U.S. Fear of nuclear annihilation was the original motivating factor, but the organiza-

**Each of these models for doing ethics at scale has opportunities and limitations.**

tion also advocated for broad ethical changes in tech research and practice, including prioritizing privacy, participatory design, and community networking.<sup>4</sup> During the Cold War, CPSR drew upon the technical expertise of its members to argue for limiting when and how computing could be used in war. CPSR members studied military technology research agendas, educated each other, and—mobilizing their expertise—publicly critiqued plans for computerized nuclear weapons. They argued that the government and military exaggerated the power of computers and identified limits to the reliability of weaponry software that could not be tested in realistic situations. To spread this message and create policy change, CPSR members distributed flyers at computing conferences, hosted meetings and speakers, studied policies and plans, gave interviews, and published analysis in their newsletters,

bulletin boards, email, reports, academic papers, books, traveling slide-shows, trade and local press, and the national media.

CPSR built a broad coalition of experts who leveraged their status to convince the public about the limits of computing technology for nuclear war. By arguing against using computers for nuclear war, CPSR members took risks that put them outside the mainstream of their field, potentially jeopardizing job opportunities and research funding. And the work of running a nonprofit was challenging: there were always financial worries, challenges attracting and managing volunteer members, and concern about keeping the organization focused on core values. Tech workers and computing researchers can change the culture of computer science through advocacy but must be willing to take personal and organizational risks.




**Ethics Work Going Forward**

Responding to computing industry crises of ethics, tech workers and researchers are uniting to develop new guidelines for responsible computing, new forms of governance, and new advocacy groups. Each has advantages and challenges, as shown in Table 2.

Computing researchers and professionals concerned with reforming their industry should join in:

- ▶ Crafting and deploying *guidelines* such as ACM’s updated computing code of ethics (see <https://bit>

**Table 2. Advantages and challenges.**

	 Guidelines	 Governance	 Advocacy
<b>Advantages</b>	<ul style="list-style-type: none"> <li>▶ Can bootstrap from existing resources</li> <li>▶ Can be tailored to technical work</li> </ul>	<ul style="list-style-type: none"> <li>▶ Influential because enforceable</li> <li>▶ Can translate best practices to regulation</li> </ul>	<ul style="list-style-type: none"> <li>▶ Can be influential and satisfying</li> <li>▶ Democratic, bottom-up motivations</li> </ul>
<b>Challenges</b>	<ul style="list-style-type: none"> <li>▶ Expensive in time and effort</li> <li>▶ Reflect the concerns of those in the room</li> <li>▶ If voluntary, can be ignored</li> </ul>	<ul style="list-style-type: none"> <li>▶ Hard to assess real vs. bureaucratic change</li> <li>▶ May not account for a diverse range of experiences</li> </ul>	<ul style="list-style-type: none"> <li>▶ Expensive in effort and resources</li> <li>▶ Outcomes can be hard to assess</li> <li>▶ Professionally risky</li> </ul>

## The work of doing computer ethics is crucial, but it is never complete.

ly/2XtN4Kq) and IEEE's recommendations for ethically aligned design (see <https://bit.ly/3tEMDbU>);

- ▶ Supporting *governance* through environmental, social, and governance (ESG) criteria, hiring of Chief Ethics Officers,<sup>5</sup> and new developing new approaches to support ethical behavior;

- ▶ Pairing *governance* and *advocacy* (for example, unionizing tech workers and researchers to influence corporations and universities);

- ▶ *Advocating* for the computing profession by hiring and supporting Black, Indigenous, and people of color in the profession, and engaging in antiracist projects;

- ▶ Establishing *governance* through new credentialing requirements in the field, such as certifications in computing, information or data ethics;

- ▶ Using design *guidelines* (for example, participatory design in UX and FACT guidelines in machine learning) that incorporate input from minoritized publics and increase transparency and accountability;

- ▶ Engaging in *advocacy* that helps the public understand the limits of computing (for example, campaigns that have resulted in restrictions of the use of AI in public spaces by government agencies and private companies;<sup>6</sup> and

- ▶ Establishing *governance* by encouraging publication venues to require explicit reflection on ethics.

As earlier models from computer security indicate, each of these models for doing ethics at scale has opportunities and limitations. And we add one last lesson from our research into ethics in computer security: these efforts depend on sustained work on

Sisyphian tasks. In studying these cases we spoke with dozens of participants; none felt that their work was complete. Many had regrets and worried that they had “dropped the ball” at some point or that their task was overwhelming. Cultural change for ethics and responsibility is slow, non-linear, and requires multiple—sometimes even competing—tactics. We worry that ethics efforts will slow as new guidelines fail to influence everyone, as new modes of governance controversially exclude some forms of innovation or overlook stakeholder groups, and advocacy groups struggle to raise funds or stay relevant as the news cycle turns.

### Conclusion

We end with a plea to persevere through the imperfectness (and sheer difficulty) of ethics work. The work of doing computer ethics is crucial, but it is never complete. Researchers and professionals—we drew our examples from CPSR members, *Menlo* participants, and conference review committees—have engaged in change despite knowing its limitations. We hope more will follow their examples. **□**

### References

1. Amrute, S. Of techno-ethics and techno-effects. *Feminist Review* 123, 1 (2019), 56–73.
2. Bruckman, A. “Have you thought about . . .”: Talking about ethical implications of research. *Commun. ACM* 63, 9 (Sept. 2020), 8–40.
3. Dittrich, D. and Kenneally, E. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*. Department of Homeland Security, 2012.
4. Finn, M. and DuPont, Q. From closed world discourse to digital utopianism: The changing face of responsible computing at Computer Professionals for Social Responsibility (1981–1992). *Internet Histories* 4, 1 (2020), 6–31.
5. Metcalf, J. et al. Owning ethics: Corporate logics, Silicon Valley, and the institutionalization of ethics. *Social Research: An International Quarterly* 86, 2 (2019), 449–476.
6. Metz, R. Portland passes broadest facial recognition ban in the US. CNN. (2020); <https://cnn.it/3zakCKG>
7. Sharma, A. Linux bans University of Minnesota for committing malicious code. *BleepingComputer* (2021); <https://bit.ly/3tHFoQD>
8. Wikipedia contributors. 2021. Sokal affair. Wikipedia, *The Free Encyclopedia*; <https://bit.ly/3EiFy5F>

**Katie Shilton** (kshilton@umd.edu) is an associate professor at the University of Maryland, College of Information Studies (iSchool), College Park, MD, USA.

**Megan Finn** (megfinn@uw.edu) is a Fellow in the Center for Advanced Study in the Behavioral Sciences at Stanford University Palo Alto, CA, USA, and an associate professor, Information School, at the University of Washington Seattle, WA, USA.

**Quinn DuPont** (quinn.dupont@ucd.ie) is an assistant professor in the School of Business at the University College Dublin, Dublin, Ireland.

Copyright held by authors.

## Coming Next Month in COMMUNICATIONS

**The Hardware Lottery**

**Datasheets for Datasets**

**AI-based Framework for Telemonitoring Heart Disease**

**Digital Agriculture for Small-Scale Producers**

**Speculation Taint Tracking**

**Software-Defined Cooking Using Microwaves**

**Declarative Machine Learning Systems**

**Digging into the Big Provenance (with SPADE)**

**What Every Engineer and Computer Scientist Should Know**

**A Q&A with Scott Aaronson**

**Plus, the latest news about augmented reality displays, trouble at the source, and the energy costs of life online.**

▶ Mark Guzdial, Column Editor

## Education Explicative Programming

*Making Computational Thinking relevant to schools.*

**W**ITH SUBSTANTIAL CONFUSION left regarding the meaning, as well as the purpose, of Computational Thinking (CT), 15 years after Jeanette Wing’s seminal *Communications Viewpoint*,<sup>7</sup> two different schools of computational thought have emerged. In the more prominent school of thought—let’s call it vocational education—the boundary between programming and CT is somewhat blurry. No doubt, the understanding of coding constructs is an essential part of CT education. However, if teaching and assessing CT is largely based on concepts such as loops, sequences, and conditional statements, then how is this fundamentally different from programming? The arguments for and against this school of thought are numerous but the most ubiquitous ones, at least in the U.S., appear to be career oriented and are grounded in predominantly economic justifications. Unfortunately, in the vocational school of computational thought the benefits of CT toward disciplines other than computer science are at best collateral.

Enter the second, less developed, school of computational thought focusing on general education. We will call this “Explicative Programming.” In this school, programming may not be in the foreground, but it becomes an interdisciplinary instrument of thought to truly



understand powerful ideas in typical K–12 disciplines such as STEM, art, language, and music. In the Explicative Programming school of computational thought, CT is about *thinking with the computer*. This idea is not new. In fact, it predates the Wing vision and can be traced back to Seymour Papert. Papert initially employed the term “Computational Thinking”<sup>3</sup> to refer to a relation between problem disciplines and what he referred to as an explicative practice of programming.

Explicative Programming facilitates CT thinking by establishing a bidirectional connection between typical K–12 disciplines and computer

science (see Figure 1). The motivational aspects of this type of connection have long been recognized by the educational reformer Dewey as “instrumental motivation.”<sup>2</sup> Instrumental motivation is an indirect kind of motivation—just like learning a few words of French may enable you to order food at a restaurant in France. In this example, instrumental motivation suggests you are mostly interested in eating food, not learning the French language. Similarly, with Explicative Programming, instrumental motivation may foster interest in either “learning to program,” or “programming to learn” or both (see Figure 1).



► **Learning to program.** Disciplines such as STEM, music, art, and languages, suggesting applications such as games, simulations, robots, stories, and animations, can serve as engaging contexts to develop programming skills.

► **Programming to learn.** Programming can serve as a dialectic to support understanding and expose misconceptions in disciplines such as STEM, music, art, and languages through hands-on inquiry processes.

It should be noted that while in theory the ideas of the vocational and explicative schools of thought could be combined, we have observed that instruction leans toward vocational education because of practical reasons. While the vision of Explicative Programming to make CT relevant to K–12 schools is highly appealing, it is incredibly challenging to implement the bidirectional connections (Figure 1, red arrows). Because of this difficulty, Explicative Programming has not yet reached systemic impact in schools. Establishing a bridge between traditional K–12 disciplines and computer science requires not only a solid understanding of both disciplines but, even more importantly, the understanding of how to meaningfully connect them. Unfortunately, teachers are quite unlikely to have relevant experience. The daunting gap between K–12 disciplines and programming needs to be connected through carefully designed bridging constructs (see Figure 1) serving as stepping stones between disciplines and programming (Figure 1, green box and arrows). Two examples employing CT patterns<sup>1</sup> as bridging constructs are described here: these projects are programmed with the AgentCubes<sup>4</sup> Computational Thinking Tool.



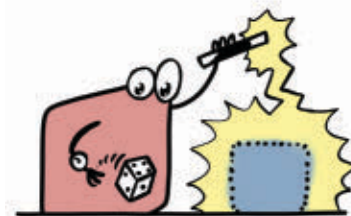
### Example 1. Collision pattern to create a virus simulation.

The virus simulation is an example of an Explicative Programming activity

## Enter the second, less developed, school of computational thought focusing on general education.

that can fit within the curriculum of a Life Science classroom. To program the virus simulation, students must program a healthy person, with some percent chance, becoming sick when coming into contact with a sick person. The bottom of Figure 2 depicts an implementation of this collision computational thinking pattern as an IF/THEN expression with a percent-chance condition of 70%. The ability to count populations of characters, allows students to experiment by changing the percent chance of becoming sick, as well as rate of death and recovery, to determine, for example, the effect different disease susceptibilities have on the population over time. This in turn enables discovery. For instance, students often try to create a deadly disease by making the death rate something close to 100%. What they find is that this, counterintuitively, leads to less

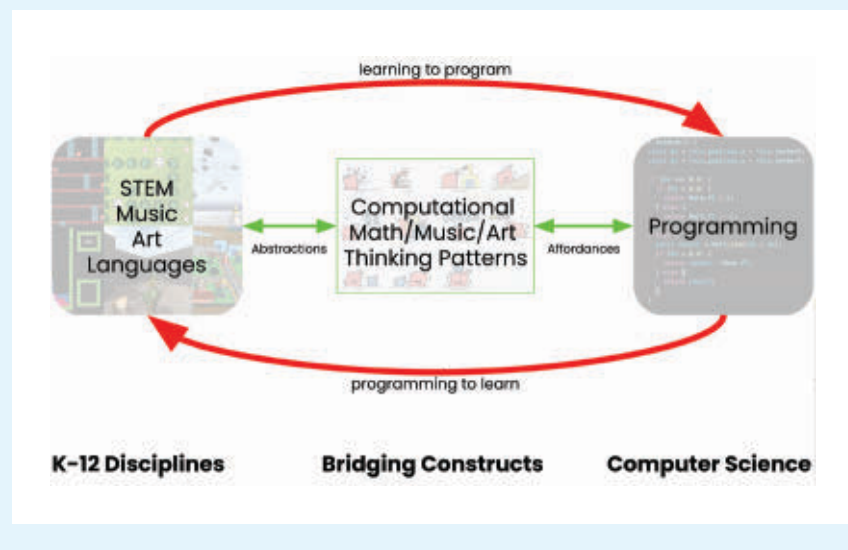
total death among the population, as many of us now know. This discovery has been witnessed from middle school to graduate school classrooms<sup>5</sup> and enrichment programs such as the June 2020 Black Girls Code Epidemiology Virtual Workshop (<https://bit.ly/BGCVirusSim>).



### Example 2. Probabilistic generation pattern to create a Frogger game.

Game design<sup>5</sup> provides ample opportunities for Explicative Programming, for example, enabling bridging constructs whereby implementation supports Computational Math Thinking. For instance, in a Frogger-like game design activity, students need to program cars driving on a busy multilane highway as an obstacle for a keyboard-controlled frog to cross. To make the game easier, but also not too simple, necessitates the correct numbers and distributions of cars on these roads. Employing the probabilistic generation pattern, students program the tunnels on the left generating cars with a certain frequency and probability. If a low frequency

Figure 1. Explicative Programming connects traditional K–12 disciplines with programming through bridging constructs.





## ACM Student Research Competition

**Attention:**  
Undergraduate *and* Graduate Computing Students

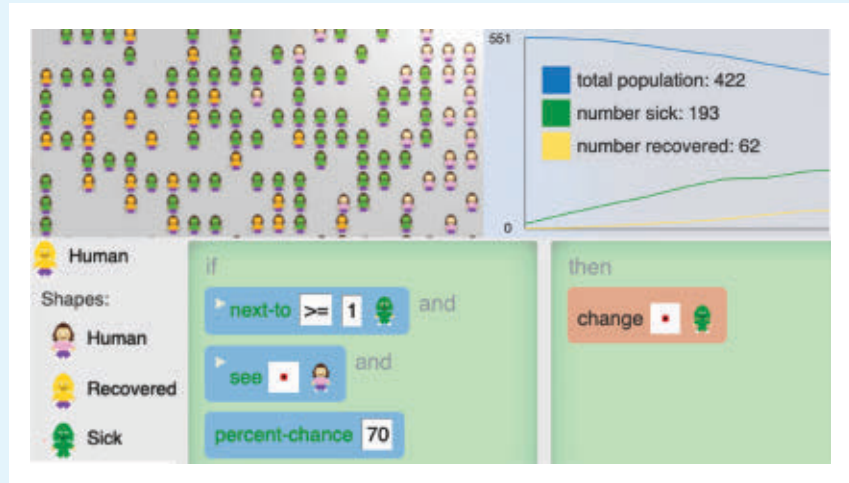
The ACM Student Research Competition (SRC), sponsored by Microsoft, offers a unique forum for undergraduate and graduate students to present their original research before a panel of judges and attendees at well-known ACM-sponsored and co-sponsored conferences. The SRC is an internationally recognized venue enabling students to earn many tangible and intangible rewards from participating:

- **Awards:** cash prizes, medals, and ACM student memberships
- **Prestige:** Grand Finalists and their advisors are invited to the Annual ACM Awards Banquet
- **Visibility:** meet with researchers in their field of interest and make important connections
- **Experience:** sharpen communication, visual, organizational, and presentation skills

Learn more:

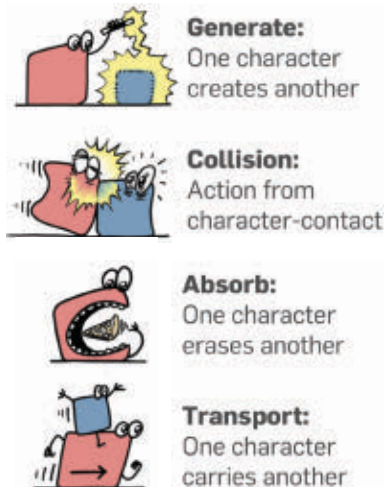
<https://src.acm.org>

**Figure 2.** The collision-change pattern implementation (bottom middle) of the epidemiology simulation in AgentCubes with population data depicted. The rule shown reads: "If I am next to one or more sick characters, and currently healthy, I become sick with a 70% chance."



(every 0.6 seconds) but high probability (100%) is used, then the result will be a wave of simultaneous cars, which looks unnatural (Figure 3, left) and makes the game play too predictable. Experimenting with the construct and selecting a high frequency (every 0.06 seconds) but low probability (10%) results in much more compelling game play (see Figure 3, right) bridging math with game design.

Our experience from teaching Scalable Game Design<sup>5</sup> to students and teachers, all over the world suggests three key principles of bridging constructs to effectively facilitate Explicative Programming.



### Bridging Constructs Should:

1) Foster creativity by serving as highly reusable concrete abstractions. Constructs should enable students to

create personally meaningful artifacts, such as games, by breaking down existing games or their own game design ideas into common design patterns they have learned to program. Just as architects are not thinking about houses in terms of nails or bricks, these patterns are at a much higher level of design abstraction than programming constructs such as IF or LOOP statements. We found Computational Thinking Patterns to be highly effective to serve as bridging constructs to build games and STEM simulations.<sup>1</sup> Computational Thinking Patterns operate at the level of phenomena describing interactions of two or more objects such as generation, collision, transport, absorption (see left), hill climbing, and diffusion. Discipline-oriented CT patterns, such as the probabilistic generate pattern employed in Example 2, are useful to engage in Computational Math Thinking, Computational Music Thinking,<sup>6</sup> and Computational Art Thinking through Explicative Programming. Programming these patterns includes the need to understand affordances of CT tools. Not all block-based, or text-based, programming languages are alike. What may be elegant and simple with one tool may be convoluted and nearly impossible with another one (<http://tiny.cc/affordances>).

2) Enable dialectic. Dialectic is a

reasoning process explaining powerful ideas by exposing misconceptions through inquiry. To make Explicative Programming dialectical the “learning to program,” and just as importantly, the “programming to learn” red arrows of Figure 1 need to be supported. Explicative Programming is not just integrated computer science education. For example, say in a science course, the teacher wants students to understand planetary motion. Students create an animation of an earth sprite moving around the sun. Earth is programmed, much like the “turtle” in Logo, using a loop to iteratively make the earth move forward some pixels and then turn right some degrees. This results in a nice animation of the earth rotating around the sun. However, it does not qualify as explicative any more than the creation of a static diorama made out of paper, wire, and tape. This activity does not engage students in a dialectical process challenging misconceptions or extending their understanding of planetary motion. In contrast, the dialectic in Example 1 is about uncovering surprising phenomena of spreading viruses, and in Example 2, the dialectic is about experimenting with compound probabilities resulting from tuning frequency and probability of multiple tunnels generating multiple cars.


3) *Facilitate instrumental motivation enabling students to create meaningful artifacts.* Instrumental motivation,<sup>2</sup> as explained previously, helps to connect disciplines with programming. Students may not be intrinsi-

## The daunting gap between K–12 disciplines and programming needs to be connected through carefully designed bridging constructs.

cally interested in programming but may want to build personally meaningful artifacts such as simulations (Example 1), games (Example 2), animations, stories, and robots. With a discipline-oriented CT pattern, such as the probabilistic generate pattern employed to create Frogger in Example 2, instrumental motivation may result in the interest to understand math/probability because its mastery helps to create a more playable game. This is in strong contrast to traditional math learning in schools where mastery of the subject is often perceived by students as hardwired aptitude and not as an enabling instrument. Unlike students, teachers may not intrinsically care about game design, but once they see students become highly engaged to learn programming

through game design teachers will likely appreciate game design as an effective pedagogic method.

### Conclusion

If Computational Thinking is to be not just another item in a crowded curriculum that must be taught, but instead wants to become truly relevant to public schools, it has to radically change its perspective toward interdisciplinary education. How can programming support the effective teaching of STEM, art, music and languages? Perhaps the generic notion of CT needs to be replaced with discipline-oriented versions such as Computational Math Thinking, Computational Music Thinking, and Computational Art Thinking. By providing educators and students constructs to bridge the vast gap between typical K–12 disciplines and programming, Explicative Programming, evolving Seymour Papert’s original conceptualization of Computational Thinking, is doing just that. 

### References

- Basawapatna, A. et al. Recognizing computational thinking patterns. Presented at the 42<sup>nd</sup> ACM Technical Symposium on Computer Science Education (SIGCSE), Dallas, TX, USA, 2011.
- Dewey, J. *Interest and Effort in Education*. Houghton Mifflin, 1913.
- Papert, S. An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning 1* (1996), 95–123.
- Repenning, A. Moving beyond syntax: Lessons from 20 years of blocks programming in AgentSheets. *Journal of Visual Languages and Sentient Systems 3* (2017), 68–89.
- Repenning, A. et al. Scalable game design: A strategy to bring systemic computer science education to schools through game design and simulation creation. *Transactions on Computing Education (TOCE) 15* (2015), 1–31.
- Repenning, A. et al. Computational music thinking patterns: Connecting music education with computer science education through the design of interactive notations. Presented at the 12<sup>th</sup> International Conference on Computer Supported Education. (Prague, 2020), 641–652.
- Wing, J.M. Computational thinking. *Commun. ACM 49*, 3 (Mar. 2006), 33–35.

**Alexander Repenning** (alexander.repenning@fhnw.ch)

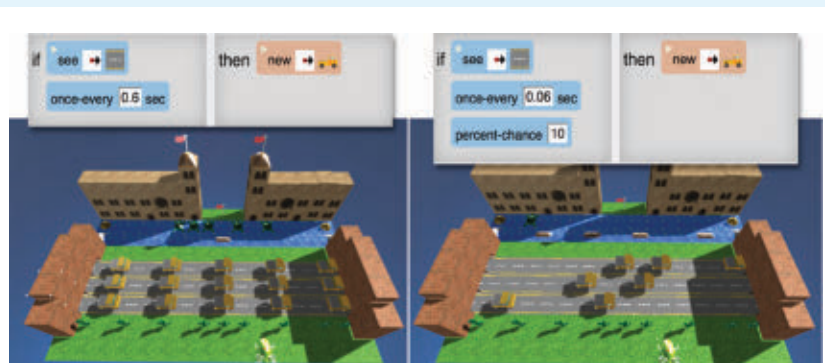
is Professor and Chair of Computer Science Education at FHNW, School of Education, Windisch, Switzerland, and a professor of Computer Science at the University of Colorado at Boulder, Boulder, CO, USA.

**Ashok Ram Basawapatna** (basawapatnaa@oldwestbury.edu) is Assistant Professor of Computer Science at SUNY Old Westbury, Department of Mathematics, Computer, and Information Science, Old Westbury, LI, NY, USA.

This research was supported by the U.S. National Science Foundation, the Swiss National Science Foundation, and the Hasler Foundation.

Copyright held by authors.

**Figure 3. Left: Generating new cars with low frequency and high probability (100%); Right: High frequency and low probability.**



## Viewpoint

# Medical Artificial Intelligence: The European Legal Perspective

*Although the European Commission proposed new legislation for the use of “high-risk artificial intelligence” earlier this year, the existing European fundamental rights framework already provides some clear guidance on the use of medical AI.*

**I**N LATE FEBRUARY 2020, the European Commission published a white paper on artificial intelligence (AI)<sup>a</sup> and an accompanying report on the safety and liability implications of AI, the Internet of Things (IoT), and robotics.<sup>b</sup> In the white paper, the Commission highlighted the “European Approach” to AI, stressing “it is vital that European AI is grounded in our values and fundamental rights such as human dignity and privacy protection.” In April 2021, the proposal of a Regulation entitled “Artificial Intelligence Act” was presented.<sup>2</sup> This Regulation shall govern the use of “high-risk” AI applications which will include most medical AI applications.

### Fundamental Rights as Legal Guidelines for Medical AI

Referring to the above-mentioned statement, this Viewpoint aims to show European fundamental rights already provide important legal (and not merely ethical) guidelines for the development and application of medical AI.<sup>7</sup>

As medical AI can affect a person’s

a See <https://bit.ly/393uYkM>

b See <https://bit.ly/3k7y11J>



physical and mental integrity in a very intense way and any malfunction could have serious consequences, it is a particularly relevant field of AI in terms of fundamental rights. In this context, it should be stressed that fundamental rights (a.k.a. human rights) not only protect individu-

als from state intervention, but also oblige the state to protect certain freedoms from interference by third parties. These so-called “*obligations to protect*” are of particular importance in medicine: For example, the European Court of Human Rights has repeatedly stated that funda-

## Fundamental rights thus constitute a binding legal framework for the use of AI in medicine.

mental rights entail an obligation on the state to regulate the provision of health services in such a way that precautions are taken against serious damage to health due to poorly provided services. On this basis, the state must, for example, oblige providers of health services to implement quality-assurance measures.

Fundamental rights thus constitute a binding legal framework for the use of AI in medicine. In line with the European motto “*United in diversity*,” this framework is distributed across various legal texts, but quite uniform regarding its content: Its core component is the European Charter of Fundamental Rights (CFR), which is applicable to the use of medical AI because the provision of medical services is covered by the freedom to provide services under European law. For its part, the CFR is strongly modeled on the European Convention on Human Rights (ECHR), which is also applicable in all E.U. states. The fundamental rights of the constitutions of the individual E.U. states also contain similar guarantees. For this reason, we focus this Viewpoint on the CFR.

### Human Oversight as a Key Criterion

It has already been emphasized by the Ethics Guidelines of the HLEG<sup>4</sup> that “European AI” must respect human dignity (Art. 1 CFR), which means medical AI must not regard humans as mere objects. From this, it can be deduced the demands for human oversight expressed in computer science<sup>1</sup> are also required by E.U. fundamental rights (see also Art. 14 of the proposed AI Act). Decisions of medical AI require human assessment before any action is taken on their ba-

sis. The E.U. has also implemented this fundamental requirement in the much-discussed provision of Art. 22 GDPR, which allows “*decisions based solely on automated processing*” only with considerable restrictions. In other words: European Medical AI legally requires human oversight (a.k.a. “a human in the loop”<sup>5</sup>).

### Explainability, Privacy by Design, and Non-Discrimination

Even more important for medical AI, however, is Art. 3 para. 2a) CFR, which requires “*free and informed consent*” of the patient. This points to a “*shared decision-making*” by doctor and patient where the patient has the ultimate say. Medical AI can therefore only be used if patients have been informed about its essential functions beforehand—admittedly in an intelligible form. This makes it clear, however, that the European fundamental rights basically require the use of explainable AI in medicine (see also Art. 13 para. 1 of the proposed AI Act).

Recent research in the medical domain<sup>6,8</sup> as well as legal research from a tort law perspective very much confirms this conclusion.<sup>3,9</sup> Consequently, European Medical AI should not be based on a “machine decision,” but much rather on “an AI supported decision, diagnostic finding or treatment proposal.” We conclude: European medical AI requires human oversight *and* explainability.

That (not only medical) European AI must be developed and operated in accordance with the requirements of protection of data and privacy (Arts. 8 and 7 CFR) and thus with the GDPR, is well acknowledged and does not require further discussion. Still, it is worth mentioning that Art. 25 GDPR not only requires controllers (“users”) of AI, but indirectly also developers of AI to take these requirements into account when designing AI applications (“privacy by design”).

There is a rich body of fundamental rights provisions requiring equality before the law and non-discrimination, including gender, children, the elderly and disabled persons in the CFR (Arts. 20–26). From these provisions, further requirements for the development and operation of European medical AI can be deduced: Not

only must training data be thoroughly checked for the presence of bias, also the ongoing operation of AI must be constantly monitored for the occurrence of bias. If medical AI is applied to certain groups of the population that were not adequately represented in the training data, the usefulness of the results must be questioned particularly critically.

At the same time, care must be taken to ensure useful medical AI can nevertheless be made available to such groups in the best possible way. In other words: European medical AI must be available for everyone. The diversity of people must always be taken into account, either in programming or in application, in order to avoid disadvantages.

### Obligation to Use Medical AI?

However, if a medical AI application meets the requirements described here, it may become necessary to explicitly impose its use for the benefit of all. European fundamental rights—above all the right to protection of life (Art. 2 CFR) and private life (Art. 7 CFR)—give rise to an obligation on the part of the state, as previously mentioned, to ensure work in health care facilities is carried out only in accordance with the respective medical due “standard of care” (a.k.a. “state of the art”). This also includes the obligation to prohibit medical treatment methods that can no longer be provided in the required quality without the involvement of AI.<sup>10</sup> This will, in the near future, probably hold true for the field of medical image processing.

### The Open Question of Liability

Does this mean the existing fundamental rights framework can answer

## European medical AI requires human oversight *and* explainability.



## Digital Threats: Research and Practice

*Digital Threats: Research and Practice* (DTRAP) is a peer-reviewed journal that targets the prevention, identification, mitigation, and elimination of digital threats. DTRAP aims to bridge the gap between academic research and industry practice. Accordingly, the journal welcomes manuscripts that address extant digital threats, rather than laboratory models of potential threats, and presents reproducible results pertaining to real-world threats.



For further information  
and to submit your  
manuscript,  
visit [dtrap.acm.org](http://dtrap.acm.org)

## The European Commission wishes to further promote the development and use of AI in Europe.

all legal questions arising for the use of medical AI? Unfortunately, there is one major exception, involving questions of liability law, which particularly unsettle the AI community. Who will be legally responsible when medical AI causes harm? The software developer, the manufacturer, the maintenance people, the IT provider, the hospital, the clinician? It is true that strict liability—liability without fault—is not unknown under European law, especially for dangerous objects or activities. Such an approach is neither required nor prohibited by the CFR, so questions of civil liability cannot be answered conclusively from a fundamental rights perspective. The European Commission is aware of this challenge and has announced in its previously mentioned report on the safety and liability implications of AI that it will evaluate the introduction of a strict liability system together with compulsory insurance for particularly hazardous AI applications—which will presumably cover most medical AI. Such a system could certainly help to eliminate many existing ambiguities regarding the liability of medical AI applications.

### Conclusion

The European Commission wishes to further promote the development and use of AI in Europe. In its white paper on AI published in 2020 it highlights the “*European Approach*” to AI, particularly referring to fundamental rights in the European Union. The authors argued, by using the example of medical AI, that many of these fundamental rights coincide with demands of computer scientists, above all human oversight, explainability and

avoidance of bias. At the same time, it is likely that medical AI will soon not only be used voluntarily, but will also have to be used by health care providers to meet the due standard of care. This makes answers to the remaining uncertainties regarding liability for defective medical AI applications more urgent. In this regard, the Commission has announced that it will soon provide clarity by proposing a strict liability approach paired with an obligatory insurance scheme for malfunctions of AI. Despite some open questions, it should nevertheless be stressed that legal requirements for the use of medical AI are already clearer today than is often assumed in computer science. **□**

### References

1. Etzioni, A. and Etzioni, O. Designing AI systems that obey our laws and values. *Commun. ACM* 59, 9 (Sept. 2016), 29–31.
2. European Commission: 'Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)'; <https://bit.ly/3AcDsCa>
3. Hacker, P. et al. Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law* 2020 28, 415–439.
4. High-Level Expert Group on Artificial Intelligence: 'Ethics Guidelines for Trustworthy AI'; <https://bit.ly/3Ad2vov>
5. Holzinger, A. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.
6. Holzinger, A. et al. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019).
7. Mueller, H. et al. The Ten Commandments of ethical medical AI. *IEEE Computer* 54, 7 (2021), 119–123.
8. O'Sullivan, S. et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *The International Journal of Medical Robotics and Computer Assisted Surgery* 15, 1 (2019), 1–12.
9. Price, W.N. Medical malpractice and black box medicine. In Cohen, G. et al. Eds., *Big Data, Health Law and Bioethics* (2018), 295–306.
10. Schönberger, D. Artificial intelligence in healthcare: A critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology* 27, 2 (2019), 171–203.

**Karl Stöger** ([karl.stoeger@univie.ac.at](mailto:karl.stoeger@univie.ac.at)) is Professor of Medical Law at the University of Vienna, Austria.

**David Schneeberger** ([david.schneeberger@univie.ac.at](mailto:david.schneeberger@univie.ac.at)) is Research Associate at the University of Vienna and currently pursuing a Ph.D. thesis on Explainable AI.

**Andreas Holzinger** ([andreas.holzinger@medunigraz.at](mailto:andreas.holzinger@medunigraz.at)) is lead of the Human-Centered AI Lab at the Medical University of Graz and currently Visiting Professor for Explainable AI at the Alberta Machine Intelligence Institute in Edmonton, Canada.

The authors are very grateful for the comments of the reviewers and the editors. Parts of this work have received funding from the Austrian Science Fund (FWF) through Project: P-32554 “A Reference Model of Explainable Artificial Intelligence for the Medical Domain.”

## Viewpoint

# We Are Not Users: Gaining Control Over New Technologies

*Seeking a more selective approach to technology usage.*

**O**N AUGUST 27, 2020, Amazon introduced its Amazon Halo: a technology comprised of AI software and a wristband that monitors body indicators including voice to detect problems, suggests a behavioral change, or other actions to potentially improve our health.<sup>a</sup> One day later, Elon Musk and his team presented their Neuralink technology—AI software and a skull chip implant that receives and sends signals to our brain to compensate for brain malfunctioning, aiming to solve various brain-related health problems.

These announcements seem like great news amid the health crisis that engulfs many of us, with technology coming to our rescue to confront some of the most critical diseases of humankind. Yet risks remain, and once the genie is out of the bottle, they are often difficult to manage and contain—they range from unintended consequences and side effects to threats to privacy and loss or misdirection of control.

Endless devices surrounding us include processors that compute and monitor our abundant but wasteful lifestyle, with generations of products getting faster, cheaper, and “better.” We cannot envision the world today without them. Further, it seems there is no escape from this trajectory, espe-



cially with the visions of smart homes, smart cities, and the like. Even without implants or providing detailed personal data to a third party, we will be constantly watched, sampled, and analyzed.

### Reflections on Technology

*“...‘modern improvements’; there is an illusion about them; there is not always a positive advance.... Our inventions are wont to be pretty toys, which distract us from serious things. They are an improved means to an unimproved end.”<sup>15</sup>*

Technology has been long known to have positive and negative effects on society. Technology critic Neil Post-

man<sup>10</sup> made the case that technology not only changes language but also our perception of ourselves. This is the nature of technology since the days of writing and then the printing press. Gutenberg’s printing press led to the revolution that eroded the central authority of the Catholic Church. Cars redefined the idea of individual freedom to move while creating suburbs, changing the identity of the collective, along with pollution and isolation that affect our health. Television, and then streaming services, created shared and then individual experiences of entertainment and information.

<sup>a</sup> See <https://bit.ly/3gap70P>

## INTERACTIONS



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

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



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

Follow us on Facebook and Twitter  

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

Association for  
Computing Machinery



## Are we progressively being made “dumb?”

This continuing trajectory of technology changes us from a collective with multiple perspectives to fragmented, less diverse “communities,” which divide communication across perspectives. Lanier<sup>5</sup> raised concerns about software technology developments leading to loss of personal identity and creativity by emphasizing the crowd, deterioration of financial soundness, and loss of the core of human free will to create, including a sense of spirituality. More recent work of Hwang<sup>4</sup> emphasizes the “Subprime attention economy” run by opaque algorithms, with its ad-driven consumption increasing the risk of a subprime mortgage-like crisis. Perrow<sup>9</sup> warned us that even the best-designed complex systems, such as nuclear or financial, will fail as unanticipated failures are embedded in them.

Thoreau's proposition<sup>15</sup> has come to haunt us again. Both the potential risks and opportunities lie all around us. Zuboff<sup>18</sup> tells us we are being surveilled all the time using the “pretty toys,” not by the state but by capitalism, which uses our data to capture our attention all the time. While Orwell<sup>7</sup> was concerned about central authoritarian structures controlling our lives, Huxley<sup>3</sup> warned us there may not be any need for a central authority if we all are made into dumb citizens. Are we progressively being made “dumb”?

### We Are Not “Dumb” Users, Are We?

Some groups have decided to tackle this head-on. Wetmore<sup>17</sup> noted the Amish are not against technology, but they decide which technology to adopt based on certain social objectives of community and cooperation. Ostrom's<sup>8</sup> work documented how communities manage their common-pool resources better locally than through central authority. Can we generalize from these examples to make communities control their destiny rather than rely on the ethics and social responsi-

bility of professionals and companies while being just users?

We must ask ourselves what we cherish: What are the technologies that make our lives better, while minimizing the risks we can anticipate? We cannot, as Thoreau<sup>15</sup> said, celebrate inventions that want to be playthings alongside inventions that matter to us as communities and individual citizens. These more useful innovations tend not to travel at warp speed, as speed is the enemy of good thinking. The way to get out of this trap is to educate each citizen to be a designer of his or her life with an understanding that each community is the place for defining what the “common good” is. Ideally, technology will develop to meet the needs of the community and that its citizens can manage, maintain, and control. It will allow time to explore and understand it while allowing to stop using and contain it.

Communities are losing the local newspapers that informed the citizens, and their leaders, about the issues of the community. They must be reconstructed and the trajectory of technology brought back to the need of the people and not the designers and their corporate sponsors. Locally sponsored community technology providers that bring the community together provide a possible solution; Townsend<sup>16</sup> recounted the need for community-controlled Internet experiments and new civics in the creation of smart cities. But the fundamental set of design questions that arise independently of technologies and context require an answer. The answer cannot be a product—yet another app—but requires a process.

Participatory processes are around in civil society, policy, and education.<sup>11</sup> To illustrate one implementation of participation that spread around the world, consider participatory budgeting, originally developed in Porto Alegre, Brazil, and its success has expanded to other policy areas, and inclusion of simulations as they have become more accessible.<sup>1,2,6</sup>

In order to facilitate successful participatory processes, we must create open spaces where diverse perspectives of all affected by a technology have a voice. The debate about a new technology should occur before it is developed



and deployed, without those providing comments being accused of impeding innovation and progress. Designers as a profession will be needed due to their artistic and technical skills, but they will cease to control the design process; they will be part of teams of professionals and intended users working together. This would require that we make everybody conscious of what design is—a true liberal art. Every citizen should be taught to understand they can contribute to design, as they would be educated in the process of design from their daily experiences, and all involved in the process would acknowledge their insights and importance in creating useful products, policies, and services.


Beyond spaces conducive to expanded participation, the design challenges we face require addressing several fundamental questions.<sup>13,14</sup> “What” and “Why” demand a careful articulation of the need that poses the challenge and the reasons underlying it. “Who” determines all those affected by a challenge and its solution, to be part of the process with those who have the skills and knowledge for addressing it. “How” addresses the means, organization, relationships, processes, or culture that will govern the process. One has to seek answers to all these questions that align together for the community and nationally. A complex challenge requires many skills and different perspectives, working together with a shared vision to create a well-rounded solution. They must be more deliberate and more informed. Paraphrasing Thomas Jefferson: “An educated citizenry is a vital requisite for our survival as a free people.” This education should include design as a liberal art to all as that will prepare us to ask the right questions that can be asked about what we need rather than our wants imagined by others. These questions are relevant for policing methods, dealing with the pandemics, or the control of our daily rituals and habits by our new wearable device. Our challenge is to ensure new technologies support us as humans, not ceding our control to them, or letting them make us dumber.

The role and understanding of design in society require constant reflection<sup>12</sup> and dialogues with those affected

## Our challenge is to ensure new technologies support us as humans, not ceding our control to them, or letting them make us dumber.

by the designs; it is the way forward to a cleaner, healthier, sustainable, safer, and saner world. We practice and teach these ideas because making them a reality requires participation and inclusion of diverse perspectives and their study and improvement. Fields of View (FOV)—a small non-profit organization—uses games and computer simulations to raise awareness of important issues through dialogue and participation in policy debates from various stakeholders including illiterate and semiliterate citizen (see <https://fieldsofview.in/>). The games FOV develops are played with stakeholders to address specific policy and operational questions, such as land use planning for a city, or solid waste collection and management, to expose the decision-making process and surface knowledge often unrevealed. The games are subsequently translated to computational simulations with data for exploration of the consequences of scenarios to enhance participation in decision making.

Another way to develop citizenry skilled in design for future effective participation is design education: we teach these ideas in design courses to hundreds of students from diverse disciplines such as engineering, architecture, stage design, social work, occupational therapy, and management, working on open-ended social and environmental challenges through the participation of and dialogue and

reflection with problem owners. The idea is to create a modular design curriculum that will cater to all students, allowing all, especially those in non-design disciplines, to develop their design skills. These implementations of spaces for participation and the design skills to join them effectively further shape our ideas and their introduction to the general public. 

### References

1. Abers, R. et al. Porto Alegre: Participatory budgeting and the challenge of sustaining transformative change. 2018.
2. Coghlan, D. and Brydon-Miller, M. *The SAGE Encyclopedia of Action Research 1-2*. SAGE Publications, 2014.
3. Huxley, A. *Brave New World*. Harper Collins, 1994.
4. Hwang, T. *Subprime Attention Crisis: Advertising and the Time Bomb at the Heart of Internet*. FSG originals (2020).
5. Lanier, J. *You Are Not a Gadget: A Manifesto*. Alfred Knopf, 2000.
6. Mouggiakou, E. et al. Participatory urban planning through online webGIS platform: Operations and tools. In *Proceedings of the 13<sup>th</sup> International Conference on Theory and Practice of Electronic Governance* (Sept. 2020), 831–834.
7. Orwell, G. *Nineteen Eighty-Four*. Secker & Warburg, 1949.
8. Ostrom, E. *Understanding Institutional Diversity*. Princeton University Press, 2009.
9. Perrow, C. *Normal Accidents: Living With High Risk Technologies—Updated Edition*. Princeton University Press, 2011.
10. Postman, N. *Technopoly: Surrender of Culture to Technology*. Vintage, 1993.
11. Reich, Y. et al. Varieties and issues of participation and design. *Design Studies* 17, 2 (1996), 165–180.
12. Reich, Y. The principle of reflexive practice. *Design Science* 3. 2017.
13. Reich, Y. and Subrahmanian, E. The PSI framework and theory of design. *IEEE Transactions on Engineering Management*. 2020.
14. Subrahmanian, E., Reich, Y., and Krishnan, S. *We Are Not Users: Dialogues, Diversity, and Design*. MIT Press, 2020.
15. Thoreau, H.D. *Walden*. Long River Press, Secaucus, NJ, 1976.
16. Townsend, A. *Smart Cities: Big Data, Civic Hackers and the Quest for New Utopia*. W.W. Norton, 2014.
17. Wetmore, J.M. Amish technology: Reinforcing values and building community. *IEEE Technology and Society Magazine* 26, 2 (2007), 10–21.
18. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for Human Future at the Frontier of Power*. Public Affairs, 2019.

**Yoram Reich** (yoramr@tauex.tau.ac.il) is a professor at the School of Mechanical Engineering and Systems Engineering Research Initiative, Tel Aviv University, Tel Aviv, Israel.

**Eswaran Subrahmanian** (sub@cs.cmu.edu) is a research professor at the Engineering Research Accelerator and Engineering and Public Policy, Carnegie Mellon University, Pittsburgh, PA, USA.

The authors thank Robin King and the anonymous reviewer for their comments that improved the exposition of the ideas presented in this Viewpoint. This Viewpoint is derived from a discussion between the authors and *Communications* Senior Editor Moshe Vardi, whose work spans many issues discussed here. Vardi worked to address unintended consequences of computing with ideas such as algorithmic verification, discussed the need to introduce laws and regulations to regulate technology, and recently, initiated a university-wide program on technology, culture, and society to address the negative impacts of technology. The ideas in this Viewpoint draw from our recent book, *We Are Not Users: Dialogues, Diversity, and Design*. MIT Press, 2020.

Copyright held by authors.

# China Region Special Section



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



# Welcome Back!

**W**ELCOME TO THE second regional special section spotlighting China. We are pleased to report there have been many exciting changes in technological advances and achievements since the first special section on China was published in the November 2018 issue of *Communications*. We hope the following articles will help you learn about and appreciate such changes.

For this special section, we launched a call for participation on January 10, 2021. This call attracted 34 proposals from 22 institutes—both academia and industry—across China. Given the pandemic situation in some cities of China, we organized a hybrid workshop on February 6, 2021, with seven sessions, including storage and blockchain, open source software and education, IoT, health informatics, accessibility, NLP and speech, AR&SLAM and vision. The workshop, which was a great success, created a live forum for people inside the region to exchange ideas and recent progress in various areas of the industry.

In an effort to expand the coverage in this section, we selected topics that were not covered in the first edition. We ultimately accepted articles covering various exciting things happening in China, including: blockchain, crowdsensing, speech, augmented reality, accessibility, system education, AI education, NLP, graph processing on supercomputers, health informatics, and the digital economy. To reflect the booming AI startups, we also invited a former journalist and now the founder of a new media startup specializing on AI-related startups and technical trends to write an article describing the AI startups working on computer vision, semiconductor, and autonomous driving. Although this issue is by no means a comprehensive view of *all* the exciting computing research in China, we believe these topics present a living snapshot of innovative activities and computing trends in China.

Finally, we would like to thank all the workshop participants and authors who contributed talks and articles, members of the *Communications* Editorial Board for their timely and extremely useful advice and guidance, and a special thanks to Long Zheng (Huazhong University of Science and Technology) for his help in organizing the workshop and assembling articles from authors. □

—*Hai Jin, Yuanchun Shi, and Dahua Lin*  
China Region Special Section Co-Organizers

**Hai Jin** is a professor at Huazhong University of Science and Technology, China.

**Yuanchun Shi** is a professor at Tsinghua University, China.

**Dahua Lin** is an associate professor in the Department of Information Engineering at the Chinese University of Hong Kong, and the Director of CUHK-SenseTime Joint Laboratory.

Copyright held by owners/authors.

## EDITORIAL BOARD

### EDITOR-IN-CHIEF

Andrew A. Chien  
eic@cacm.acm.org

### DEPUTY TO THE EDITOR-IN-CHIEF

Morgan Denlow  
cacm.deputy.to.eic@gmail.com

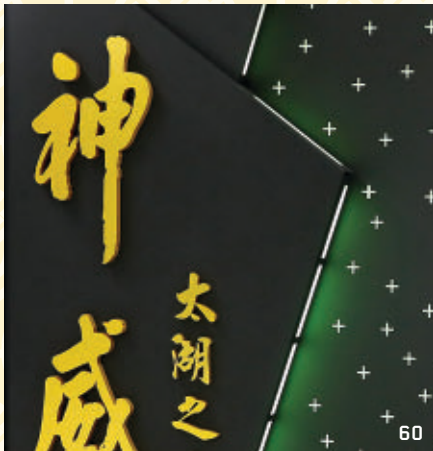
### CO-CHAIRS, REGIONAL SPECIAL SECTIONS

Jakob Rehof  
Haibo Chen  
P. J. Narayanan

### SPECIAL SECTION CO-ORGANIZERS

Hai Jin  
Huazhong University of Science and Technology, China  
Yuanchun Shi  
Tsinghua University, China  
Dahua Lin  
Chinese University of Hong Kong

## Hot Topics

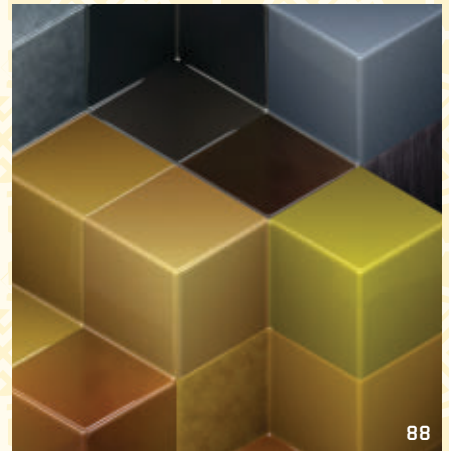


- 44 **Innovations and Trends in China's Digital Economy**  
By Fang Su, Xiao-Peng An, and Ji-Ye Mao
- 48 **Teaching Undergraduates to Build Real Computer Systems**  
By Chunfeng Yuan, Xiaopeng Gao, Yu Chen, and Yungang Bao
- 50 **Knowledgeable Machine Learning for Natural Language Processing**  
By Xu Han, Zhengyan Zhang, and Zhiyuan Liu
- 52 **AI+X Micro-Program Fosters Interdisciplinary Skills in China**  
By Fei Wu, Qinming He, and Chao Wu
- 55 **AI Start-Ups in China**  
By Jing Yang
- 57 **Natural Interactive Techniques for the Detection and Assessment of Neurological Diseases**  
By Feng Tian, Yuntao Wang, and Yicheng Zhu
- 60 **Processing Extreme-Scale Graphs on China's Supercomputers**  
By Yiming Zhang, Kai Lu, and Wenguang Chen

## Big Trends



- 64 **The Present and Future of Mixed Reality in China**  
By Guofeng Zhang, Xiaowei Zhou, Feng Tian, Hongbin Zha, Yongtian Wang, and Hujun Bao
- 70 **The Practice of Applying AI to Benefit Visually Impaired People in China**  
By Chun Yu and Jiajun Bu
- 76 **Crowdsensing 2.0**  
By Zhiwen Yu, Huadong Ma, Bin Guo, and Zheng Yang



- 81 **The Practice of Speech and Language Processing in China**  
By Jia Jia, Wei Chen, Kai Yu, Xiaodong He, Jun Du, and Heung-Yeung Shum
- 88 **Blockchain in China**  
By Liang Cai, Yi Sun, Zibin Zheng, Jiang Xiao, and Weiwei Qiu

Digital Commerce | DOI:10.1145/3481604

# Innovations and Trends in China's Digital Economy

BY FANG SU, XIAO-PENG AN, AND JI-YE MAO

**C**HINA IS BOTH A global leader in e-commerce and the world's manufacturing powerhouse.

And yet the development of an industrial Internet is far behind the booming consumer Internet, which creates a tremendous divide between consumer demand and the supply side. A distinctive feature of the recent surge of digital economy in China is the strong push from the rising consumer demand for quality products, which has a major impact on the industrial digitization.<sup>4</sup>

China has over 1.3 billion mobile Internet users, the largest online shopper population, the largest amount and highest ratio of mobile payment in the world.<sup>1</sup> According to China Statistical Yearbook (CSY) 2020, 25% of the national retail took place online in 2019, amounting to \$1.8 trillion,<sup>3,11</sup> over 90% of which was via mobile



Saturnbird Coffee's colorful capsules.

payment.<sup>11</sup> These forces have given rise to Internet giants such as Alibaba, Tencent, JD.com, and Xiaomi, which have not only shown leadership in global digital innovation<sup>8</sup> but also created a digital ecosystem in China. For example, the digital economy based on the Alibaba ecosystem accounted for over 69 million jobs.<sup>12</sup> As

a result, the digital economy in China has been growing rapidly, at an annual rate of 16.6%<sup>13</sup> between 2015 to 2020, the fastest in the world, and its size, ranked second in the world behind the U.S.,<sup>4</sup> exceeded \$6.07 trillion or 38.6% of the GDP in 2020.<sup>6</sup>

Though China has been the largest manufacturer in

the world for over a decade, the majority of firms fall behind in digital transformation. One-third of them do not have a website and only 10.2% manufacturers conduct online transactions, according to CSY 2020. Moreover, industrial digital economy accounts for only 19% of industrial GDP,<sup>5</sup> which is below the global

average 23.5% and far lower than Germany's 45.3%, the highest in the world.<sup>4</sup> An Accenture report on digital transformation shows that digital technologies did not yield a satisfactory return in two-thirds of the firms in China.<sup>9</sup> Therefore, the manufacturing industry has become the primary battlefield in the era of digital economy in China.<sup>10</sup>

Meanwhile, millennials (those born between 1980–1994) and Generation Z (born between 1995–2009) account for nearly 40% of the Chinese population, who are indigenous netizens representing the most powerful force in the Chinese consumer market.<sup>7</sup> The mobile Internet has become an indispensable part of their daily life. The new generations of consumers exhibit a strong dependence on smartphones and social media, reliance on key opinion consumers (KOCs) and key opinion leaders (KOLs) in making purchasing decisions, pride in domestic brands, preference for smart products and visual attractiveness, attention to product attributes related to health, entertainment, and experience, and willingness to pay extra for high quality. Their attitudes toward digital technologies, lifestyles, and shopping experiences create both opportunities and challenges for traditional firms.

It is against such a backdrop that a new breed of digital start-ups has emerged, and some of them have rapidly evolved into a unicorn in its industry, by bridging the demand and supply sides with the mobile Internet. This emerging trend could represent the future of digital economy. This article examines how the fusion of the world's largest consumer Internet and manufacturing powerhouse has created this

new breed of digital firms in the highly competitive red ocean of consumer product industries, in terms of the critical success factors, and what lessons can be learned by other parts of the world.

### Critical Success Factors for the Digital Start-ups

This article examines three successful digital start-ups, which share many common characteristics in becoming the top domestic brand in their respective industry (see accompanying table) in three to four years after foundation, in terms of five critical success factors.

The first factor is product innovation that fulfills a consumer need and relieves their pain with traditional products. For example, Saturnbird Coffee produces quality instant coffee that offers the taste and healthy ingredients of fresh coffee and yet the convenience of instant coffee together. Similarly, YQSL provides “healthy and tasty” beverages, which are in short supply on the market. Perfect Diary addresses consumers' headaches with expensive international brands of cosmetics, and their desire for affordable domestic brands for young women in the 18–28 age group.

Moreover, Saturnbird Coffee designed capsules in bright yellow, red, and brown colors, with an attractive visual identity to appeal to the younger generation of consumers, who prefer distinctive and fun packaging. YQSL paints a big Japanese character “气” on beverage bottles to be conspicuous among competing brands. Perfect Diary's highly popular lipsticks take the shape of a narrow high heel of ladies' shoe.

The second and the most distinctive factor is consum-

## New digital start-ups collect and use big data extensively to gain insights to consumer behaviors and to optimize interactions with consumers.

er-centric operation based on the Internet, which is also the fundamental difference between these digital start-ups and traditional firms. Starting from R&D, every phase of the operation involves extensive interactions with consumers. KOCs are invited for trial use and feedback in an iterative manner, till the product becomes satisfactory. Subsequently, a wide range of consumers are invited for product sampling, and the iteration could repeat several rounds before product releases. During the sales and adoption phases, the digital start-ups encourage content-generation by consumers to share their experiences on the Internet. Such end-to-end interaction with consumers allows precise and timely insights to consumer needs, while increasing brand

recognition and emotional attachment. Moreover, not only do the extensive interactions with consumers shorten R&D cycle time, but also allow frequent releases of new products. For example, Perfect Diary rolled out five to six new designs per month on average, releasing over 1,500 new SKUs (Stock Keeping Unit, corresponding to a scannable bar code for inventory management) during 2019 and 2020. The typical cycle time is less than six months from idea conception to product release, much shorter than the 7–18 months required by international brands.

Attractive product packaging draws the attention of KOCs and consumers and prompts the sharing of photos and stories. In fact, over 90% of the content about Saturnbird Coffee on a popu-

### Profile of the three focal cases.

	Saturnbird Coffee	Yuan Qi Sen Lin (YQSL)	Perfect Diary
<b>Product Category</b>	Quality instant coffee	Sugar-free beverages	Cosmetics
<b>Foundation</b>	2015	2016	2017
<b>Ranking by sales in its category</b>	No. 1 coffee brand (ahead of Nestle) in June 18 Shopping Festival, 2019, and Nov. 11 Shopping Festival, 2020	No. 1 beverage brand (ahead of Coco-Cola) during June 18 and Nov. 11 Shopping Festivals, 2020	No. 1 beauty makeup brand (ahead of L'Oréal) during Nov. 11 Shopping Festival, 2020
<b>Revenue</b>	\$23 million in 2019	\$450 million in 2020	\$570 million in 2020
<b>Market Value</b>	Raised over \$15 million in series B financing	\$2 billion estimated	\$7 billion (stock price on NYSE)

lar social media platform is user-generated, which draws additional consumers. Information on popular products and promotional activities is also disseminated via the Internet in the forms of product sampling, seminars and short video tips, and infomercials.

Consumer engagement is established to maximize coverage via all channels including major online marketplaces, WeChat (similar to Facebook), social commerce apps, Weibo (Twitter equivalent), and online forums. For example, Perfect Diary uses a WeChat personal account to post promotion materials on its moments first, and then broadcasts details to subscribers of the firm’s official WeChat account, and lastly disseminates promotions and URLs for purchasing in WeChat groups, in a closed circle. The timing of each of these steps is based on big data on consumers’ daily Internet use habits, to maximize exposure and impact.

The third factor is big data-driven decision-making. The new digital start-ups collect and use big data extensively to

gain insights to consumer behaviors and to optimize interactions with consumers. For example, Perfect Diary established an IT and data team, which accounted for 20% of the headcounts at the firm’s headquarters. The team built up the IT infrastructure, consumer big data, a social media-based marketing engine, and a consumer interaction platform. Saturnbird Coffee made use of business analytics from Tmall (the leading business-to-consumer online marketplace for brand name stores to sell products to consumers), which revealed that a consumption habit would be established if someone drinks 50 capsules of coffee in two consecutive months followed by additional purchase in 90 days. In response, the previous nine-capsule package was replaced with the 24-capsule one, and incentives were given for purchasing two packages together. Furthermore, a 64-capsule package was launched later, which became a best-seller.

Fourth, contract manufacturing was used along with collaborative R&D.

For example, Saturnbird Coffee collaborated with a manufacturer in experimenting with cold extraction and freeze-dried powder techniques used in other industries and adapted it to the making of instant coffee. The result was pour-over coffee that could melt in water of any temperature in just three seconds. YQSL and Perfect Diary collaborated with contract manufacturers for international brands of similar products, with capabilities and experience for technological innovation.

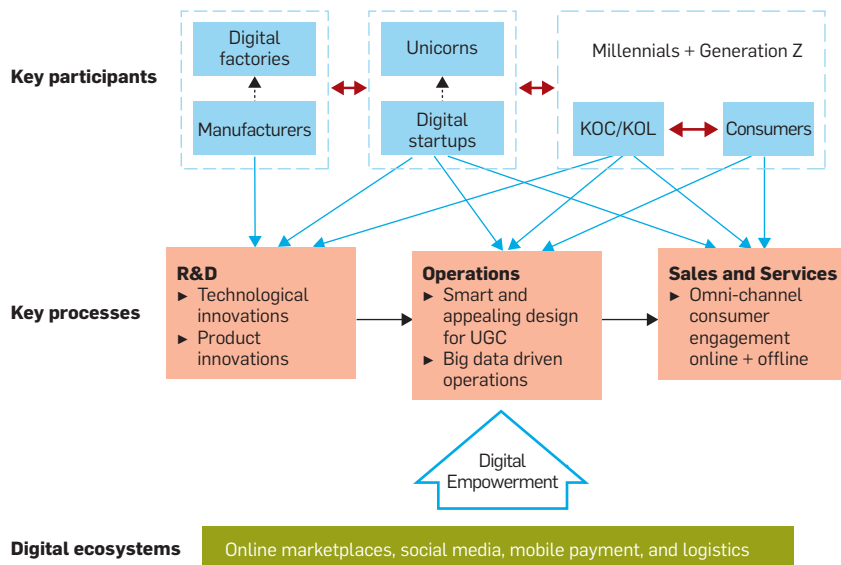
This collaboration model entails a low start-up cost, drawing upon the strength of China’s manufacturing industries. Working with established manufacturers provides not only quality guarantee, but also flexibility in adjusting production techniques and new supplementary ingredients for rapid iterations and scaling-up. However, after significant growth in brand power, sales, and financing, the digital start-ups often chose to build their own factory and R&D center for stronger control over technological innovation and processes.

This could result in more fundamental reshaping of the industry.

Fifth and lastly, the digital ecosystem played an empowerment role, including facilitating international expansion and providing Internet talents, in addition to big data and business analytics. Soon after product launching, these digital start-ups started exporting via existing online platforms, contrary to the practices of earlier generations of Internet firms such as Alibaba and Tencent, which entered overseas market only after achieving dominance in China. For example, Perfect Diary launched its website for overseas consumers at its third anniversary, by establishing its storefront on Tmall & Taobao Overseas e-commerce website. Similarly, YQSL and Saturnbird Coffee also took the same route. YQSL beverages are sold in over 30 countries and regions including the U.S., Japan, and Canada. Furthermore, each of these digital start-ups attracted much needed talents from e-commerce firms, and thus the founder team involved a combination of talents from the traditional businesses and e-commerce firms.

**A Digital Economy Driven by the Consumer Internet and Empowered by a Digital Ecosystem**

The success of the three digital start-ups demonstrates a winning formula for digital economy (see the accompanying figure), which is to capture consumer needs via the Internet, contract manufacturing along with collaborative product innovation, and extensive consumer participation in R&D and marketing. The consumer Internet is pulling the industrial Internet for-



The ecosystems for digital start-ups.



ward, which narrows down the huge gap between their different stages of development. In general, the fusion between the flourishing e-commerce and the robust manufacturing powerhouse has created fertile soil for the rapid growth of digital economy in China.

A distinctive characteristic of the current wave of digital economy in China is that it is driven by strong emerging consumer demand, which pushes technological and business model innovation.<sup>4</sup> This driving force is quite different from that in the developed world, for example, technological innovation in the U.S.,<sup>4</sup> and advanced industrial bases in western Europe, Japan, and South Korea. Moreover, Chinese digital start-ups are consumer-centric, and their operation heavily depends on the Internet, featuring end-to-end consumer interactions from conception of product idea to consumer experience after-sales and adoption. Such a process helps create not only best-seller products but also consumer loyalty.

This research offers two key insights for other parts of the world for reference. First, it is important to leverage the consumer Internet and ecosystems, which enable digital start-ups in many ways such as providing channels for engaging consumers directly, gaining big data on consumer behaviors, and facilitating expansion overseas. Connectivity of the Internet makes it feasible for novel products to gain consumer recognition in a much shorter time period than before. Moreover, online marketplaces offer not only business analytics but also talents who understand how the Internet and social media can be used to operate in

a consumer centric manner. This growth pattern of digital economy might hold true for other parts of the world, that is, a well-established consumer Internet and digital ecosystem play a key role in enabling digital start-ups. Second, in keeping pace with the digital economy, there will be a surge in the demand for digital talents needed not only for developing the IT infrastructure, natural language processing tools, and analytics platforms, but also acquiring and analyzing user-generated contents and consumer behavioral data on the Internet. The computer science and related fields would be a primary source for providing such digital talent and should be prepared for this new demand.

### Future Challenges and Directions

The three digital start-ups analyzed in this article demonstrate exciting new trends in the digital economy in China, which represents a fusion with the traditional economy, and efficient bridging between consumer demand and the supply. These new trends could also be instrumental for integrating the Chinese economy into the global value chain and supply chain. Nevertheless, the digital economy in China faces serious challenges ahead.

First, there is a huge disparity in regional development of digital economy. Digital economy accounts for over 40% of the GDP in the more developed eastern coastal regions, compared to under 25% in the northwestern regions with insufficient innovation and resources. Given the differences in regional industry structures and resource endowment, a universal development path for digital economy does not

## These new trends could be instrumental for integrating the Chinese economy into the global value chain and supply chain.

exist. Therefore, a key challenge is how to promote the fusion between traditional economy and digital economy along with cross-regional collaboration.

Second, the disparate distribution of digital infrastructures and human talents is another major hindrance for digital transformation in the majority of Chinese firms, which creates a new digital divide. The best digital talents and advanced digital infrastructures are concentrated in Internet companies, leaving few for the traditional economy. As a result, many traditional firms are unable to take advantage of the consumer Internet, despite the existence of digital ecosystems including social media, online marketplaces, and online payment systems. Therefore, a national priority remains to upgrade the IT infrastructure to enable digital innovation, and to educate digital talents.

Third, as demonstrated earlier in this article, digital start-ups must expand overseas for greater success and economies of scale, while maintaining a leading position in the domestic market. To this end, they should pay close attention to global product standards and technological innovation and stay committed to R&D collaboration with overseas partners for mutual benefits. 

### Further Reading

1. Accenture. Report on Chinese consumer insights (in Chinese), 2018; <http://www.199it.com/archives/732743.html>
2. Accenture. Research on indices of digital transformation in Chinese enterprises (in Chinese), 2020; <http://www.199it.com/archives/1131083.html>
3. AliResearch. Report on the development of Chinese consumer brands (in Chinese), 2020; <https://bit.ly/3gPdFso>
4. CAICT. Profile of global digital economy (in Chinese), 2020.
5. CAICT. White paper on the development of Chinese digital economy (in Chinese), 2020.
6. CAICT. White paper on the development of Chinese digital economy (in Chinese), 2021.
7. Euromonitor International. How China's Urban Millennials and Generation Z live and spend, 2020; <https://www.baogaoting.com/info/19115>
8. Herrero, A.G., Xu, J. How big is China's digital economy? Bruegel Working Paper, 2018/04, Bruegel, Brussels.
9. IResearch. Report on the path for digital transformation practices (in Chinese), 2020; <http://finance.sina.com.cn/jjxw/2021-02-23/doc-ikftssap8326521.shtml>
10. Li, K., Kim, D.J., Lang K.R., Kauffman, R.J., and Naldi, M. How should we understand the digital economy in Asia? Critical assessment and research agenda. *E-Commerce Research and Applications* 44 101004 (2020).
11. Ministry of Commerce, People's Republic of China. 2019 Report on the development of national online retail market (in Chinese); <https://dzswgf.mofcom.gov.cn/news/5/2020/4/1586913870177.html>
12. School of Labor and Human Resource, Renmin University of China. A report on the job structure and job quality in the Ali ecosystem (in Chinese), 2020; <https://bit.ly/3vLG1be>
13. Yang, Y. National digital economy grew over 16.6% during the 13<sup>th</sup> five-year plan period (in Chinese), 2021; <https://bit.ly/3zRWCXu>

**Fang Su** is an associate professor at the Management School of Jinan University, Guangzhou, China.

**Xiao-Peng An** is vice president of Ali Research Institute, Beijing, China.

**Ji-Ye Mao** is a professor and dean of the Business School of Renmin University of China, Beijing.

© 2021 ACM 0001-0782/21/11

# Teaching Undergraduates to Build Real Computer Systems

BY CHUNFENG YUAN, XIAOPENG GAO, YU CHEN, AND YUNGANG BAO

**C**OMPUTER SYSTEM COURSES (for example, computer organization, computer architecture, operating system, and compiler) are the foundation of computer science education. However, it is difficult for undergraduates to fully grasp key concepts and principles of computer systems due to the gap between theory and practice. To mitigate the gap, Chinese educators have spent the last decade focusing on teaching undergraduates to build real computer systems. They carried out many effective reform measures with the philosophy of learning-by-doing, which have significantly improved the computer system skills and abilities of Chinese undergraduates.

*Computer system education in China before 2010.* In 2013, Yuan and



Students at Tsinghua University in Beijing.

Chen conducted a study that analyzed the scores of computer system courses in the National Entrance Examination to Graduate School for the years of 2009–2012. The statistical results show the average scores were extremely low, ranging from only 34.3 to 49.3 during the four years and the scores of operating systems were a bit better, ranging from 41.3 to 54.4.<sup>3</sup> This situation stunned

Chinese educators for it revealed serious problems in computer science education before 2010 in China.

Specifically, there were two main problems: First, computer science courses mainly focused on memorizing concepts rather than practice training, therefore most undergraduates never dived into challenging tasks such as building a CPU or an OS kernel from scratch. Second, knowledge was taught in a disconnected manner, so that undergraduates lacked the overview of a whole computer system with multiple layers, such as programs, compilers, operating systems, instruction set architectures (ISAs), and computer architecture. When encountering a problem requiring knowledge of multiple layers, students usually became confused without knowing how to start.

## Bridging the Gap between Theory and Practice

Recognizing the problems, Chinese educators decided to reform computer system education under the leadership of the Steering Committee on Computer Education of the Ministry of Education. Eight universities pioneered computer system education reform, aimed at determining curricula that not only covered basic concepts and key principles but also emphasized applying knowledge of multiple layers to solve problems. Educators of these universities rewrote textbooks to organize computer system knowledge with a more comprehensive structure, redesigned course projects with sufficient challenging tasks and built teaching platforms to facilitate students' practice. Promoting students' system ability soon became one of the

**Chinese educators have devoted exhaustive efforts over the past 10 years to reform measures for improving the technical skills of undergraduates by teaching them to build real computer systems.**

hottest topics in the computer education field in China and more than 100 universities joined together to reform their curriculums. Educators also found that competition is an effective measure for improving computing abilities and then launched several nationwide competitions.

### The Practice of Learning-by-Doing

Chinese educators have devoted exhaustive efforts over the past 10 years to reform measures for improving the technical skills of undergraduates by teaching them to build real computer systems. Generally, there are four kinds of practices:

► *Project assignments:* Several universities, such as Nanjing University and Peking University, devised new courses that comprehensively introduce computer systems and provide challenging project assignments (PAs). For example, Nanjing University's PAs require undergraduates to complete a full system emulator step by step, which consists of a CPU and an OS and is designed to run a popular computer game.

► *FPGA-based systems:* Many universities instructed

undergraduates to build computer systems from scratch based on FPGAs. In Tsinghua University, for example, the final project of a computer organization course is to build a reduced CPU from scratch that can run applications on an FPGA board in just three weeks. During the Fall 2020 semester, 242 undergraduates took the course and successfully completed the final project. Meanwhile, the OS course project and the network course project involve building an operating system and a router from scratch respectively.

► *Real chips:* Some universities instructed undergraduates to build real chips. In 2019, University of Chinese Academy of Sciences launched the One Student One Chip (OSOC) Initiative. Five undergraduates participated in the OSOC Initiative and completed a 64-bit RISC-V processor SoC named NutShell,<sup>2</sup> which was manufactured by SIMC 110nm technology and can successfully run Linux at 200MHz. In 2020, 11 undergraduates from five universities participated in the second-year OSOC Initiative and built nine different RISC-V CPU cores that were tapped out

## Competitions have significantly promoted the active atmosphere of computer system education and improved the computing capabilities of undergraduate students in China.

in December 2020.

► *Competitions:* In 2017, the Steering Committee on Computer Education on Computer Education launched a nationwide competition in computer system ability that has already attracted more than 2,000 undergraduates from about 100 universities to participate. The competition initially focused on CPU design and as of 2021 includes operating system topics and compiler topics, in which 201 teams and 99 teams registered to participate respectively. Over the past five years, the competition has continuously improved undergraduates' computer system skills.

In 2017, 30% of the teams were able to complete CPU designs with caches and 10% of CPU designs can run OSs, among which only one team successfully ported Linux on their CPU. By contrast, in 2019, 100% of teams completed CPU designs with caches and 54% of CPUs can run OSes. One team even completed a 4-issue out-of-order CPU design that can run on an FPGA. The competitions have significantly promoted the active atmosphere of computer system education and improved student skillsets.

the computer systems arena is still far behind developed countries. A survey of 10 years' worth of papers from ISCA/OSDI/SOSP conferences (2008–2017) showed the number of China-based first authors as only 1/20 the number from the U.S.<sup>1</sup> China has a huge demand for computing talent, however, computer science education is challenged by the need to include extensive practice training. Over the past decade, China's educators have realized the importance of practice and conducted a series of reform measures to change the status. ■

### References

1. Bao, Y., Sun, N., Zhang, K. Thinking and practices of open source design of processor chip and agile development methods. *Commun. CCF* (2019).
2. Wang, H., Zhang, Z., Zhang, L., Jin, Y., Wang, K. NutShell: A Linux-Compatible RISC-V Processor Designed by Undergraduates. *RISC-V Global Forum, 2020*; <https://github.com/OSCPU/NutShell>.
3. Yuan, C., Chen, R. The Teaching Problems of Undergraduate Computer Basic Courses from the Entrance Examination to Graduate School. *China Examination, 2013*.

**Chunfeng Yuan**, Nanjing University, Jiangsu, China.

**Xiaopeng Gao**, Beihang University, Beijing, China.

**Yu Chen**, Tsinghua University, Beijing, China.

**Yungang Bao**, University of the Chinese Academy of Sciences, Beijing, China.

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



The 64-bit NutShell RISC-V processor designed by students from the University of the Chinese Academy of Sciences.

### Conclusion

China's pool of talent in

# Knowledgeable Machine Learning for Natural Language Processing

BY XU HAN, ZHENGYAN ZHANG, AND ZHIYUAN LIU

**I**N THE PAST decades, one line has run through the entire research spectrum of natural language processing (NLP)—*knowledge*. With various kinds of knowledge, such as linguistic knowledge, world knowledge, and commonsense knowledge, machines can understand complex semantics at different levels. In this article, we introduce a framework named “knowledgeable machine learning” to revisit existing efforts to incorporate knowledge in

NLP, especially the recent breakthroughs in the Chinese NLP community.

Since knowledge is closely related to human languages, the ability to capture and utilize knowledge is crucial to make machines understand languages. As shown in the accompanying figure, the symbolic knowledge formalized by human beings was widely used by NLP researchers before 1990, such as applying grammar rules for linguistic theories<sup>3</sup> and building knowledge bases for expert systems.<sup>1</sup> After 1990, statistical

learning and deep learning methods have been widely explored in NLP, where knowledge is automatically captured from data and implicitly stored in model parameters. The success of the recent pretrained language models (PLMs)<sup>4,13</sup> on a series of NLP tasks proves the effectiveness of this implicit knowledge in models. Making full use of knowledge, including both human-friendly symbolic knowledge and machine-friendly model knowledge, is essential for a better understanding of languages, which has gradually

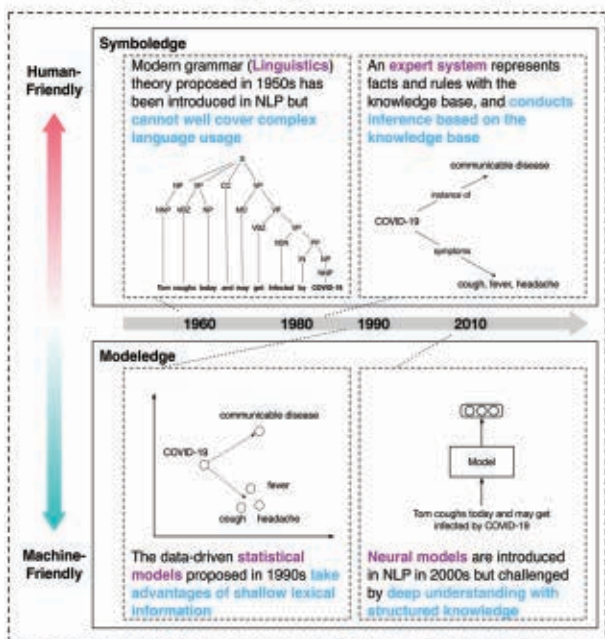
become the consensus of NLP researchers.

The spectrum depicted in the figure shows how knowledge was used for machine language understanding in different historical periods. The framework shows how to inject knowledge into different parts of machine learning.

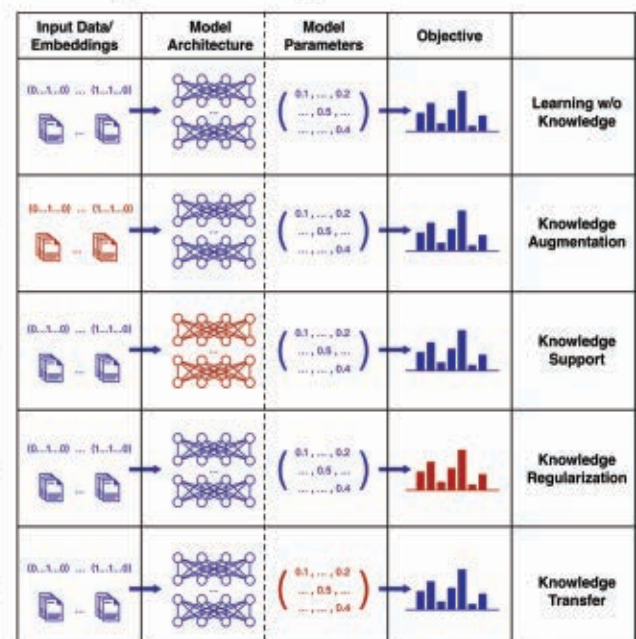
## Knowledgeable ML for NLP

To clearly show how to utilize knowledge for NLP tasks, we introduce knowledgeable machine learning. Machine learning consists

### Knowledge for Language Understanding



### Knowledgeable Learning for NLP



A historical glimpse of the NLP research spectrum and the whole framework of knowledgeable machine learning.

of four components: input, model, objective, and parameter. As shown in the figure, knowledgeable machine learning aims at covering the methods that apply knowledge to enhance these four machine learning components. According to which component is enhanced by knowledge, we can divide existing methods utilizing knowledge for NLP tasks into four categories:

**Knowledge augmentation** enhances the input of models with knowledge. There are two mainstream approaches for knowledge augmentation: one is to directly add knowledge into the input, and the other is to design special modules to fuse the original input and related knowledgeable input embeddings. So far, knowledge augmentation has achieved promising results on various tasks, such as information retrieval,<sup>11,18</sup> question answering,<sup>10,15</sup> and reading comprehension.<sup>5,12</sup>

**Knowledge support** aims to bolster the processing procedure of models with knowledge. On one hand, knowledgeable layers can be used at the bottom for preprocessing input features, and features can thus become more informative, for example, using knowledge memory modules<sup>6</sup> to inject informative memorized features. On the other hand, knowledge can serve as an expert at top layers for post-processing to calculate more accurate and effective outputs, such as improving language generation with knowledge bases.<sup>7</sup>

**Knowledge regularization** aims to enhance objective functions with knowledge. One is to build extra objectives and regularization functions. For example, distantly supervised

learning utilizes knowledge to heuristically annotate corpora as new objectives and is widely used for a series of NLP tasks such as relation extraction,<sup>8</sup> entity typing,<sup>17</sup> and word disambiguation.<sup>9</sup> The other approach is to use knowledge to build extra predictive targets, such as ERNIE,<sup>20</sup> CoLAKE,<sup>14</sup> and KEPLER,<sup>16</sup> which take knowledge bases to build extra pre-training objectives for language modeling.


**Knowledge transfer** aims to obtain a knowledgeable hypothesis space and make it easier to achieve effective models. Both transfer learning and self-supervised learning focus on transferring knowledge from labeled and unlabeled data respectively. As a typical paradigm of transferring model knowledge, fine-tuning PLMs has shown promising results on almost all NLP tasks. Some Chinese PLMs like CPM<sup>21</sup> and PanGu-alpha<sup>19</sup> have recently been proposed and have shown awesome performance on Chinese NLP tasks. CKB<sup>2</sup> has further been proposed to build a universal continuous knowledge base to store and transfer model knowledge from various neural networks trained for different tasks.

Besides the studies mentioned here, many researchers in the Chinese NLP community are committed to using knowledge to enhance NLP models. We believe all these efforts will advance the development of NLP toward better language understanding.

## Conclusion

In this article, we introduced a knowledgeable machine learning framework to show existing efforts of

# Since knowledge is closely related to human languages, the ability to capture and utilize knowledge is crucial to make machines understand languages.

utilizing knowledge for language understanding, especially some typical works in the Chinese NLP community. We hope this framework can inspire more efforts to use knowledge for better language understanding. 

## References

- Avron, B. and Feigenbaum, E.A. *The Handbook of Artificial Intelligence*. 1981.
- Chen, G., Sun, M. and Liu, Y. Towards a universal continuous knowledge base. 2020; arXiv:2012.13568.
- Chomsky, N. *Syntactic Structures*. De Gruyter, 1957.
- Devlin, J., Chang, M-W, Lee, K. and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL'2019*, 4171–4186.
- Ding, M., Zhou, C., Chen, Q., Yang, H. and Tang, J. Cognitive graph for multi-hop reading comprehension at scale. In *Proceedings of ACL'2019*, 2694–2703.
- Ding, M., Zhou, C., Yang, H. and Tang, J. CogLTX: Applying BERT to long texts. In *Proceedings of NeurIPS'2020*, 12792–12804.
- Gu, Y., Yan, J., Zhu, H., Liu, Z., Xie, R., Sun, M., Lin, F. and Lin, L. Language modeling with sparse product of sememe experts. In *Proceedings of EMNLP'2018*, 4642–4651.
- Han, X., Liu, Z. and Sun, M. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of AAAI'2018*, 4832–4839.
- Huang, L., Sun, C., Qiu, X. and Huang, X-J. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of EMNLP-IJCNLP'2019*, 3500–3505.
- Liu, K., Zhao, J., He, S. and Zhang, Y. Question answering over knowledge bases. *IEEE Intelligent Systems* 30, 5 (2015), 26–35.
- Liu, Z., Xiong, C., Sun, M. and Liu, Z. Entity-Duet Neural Ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of ACL'2018*, 2395–2405.
- Qiu, D., Zhang, Y., Feng, X., Liao, X., Jiang, W., Lyu, Y., Liu, K. and Zhao, J. Machine reading comprehension using structural knowledge graph-aware network. In *Proceedings of EMNLP-IJCNLP'2019*, 5898–5903.
- Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. Improving language understanding by generative pre-training. OpenAI Blog, 2018.
- Sun, T., Shao, Y., Qiu, X., Guo, Q., Hu, Y., Huang, X-J and Zhang, Z. CoLAKE: Contextualized language and knowledge embedding. In *Proceedings of COLING'2020*, 3660–3670.
- Wang, L., Zhang, Y., and Liu, T. A deep learning approach for question answering over knowledge base. *Natural Language Understanding and Intelligent Applications*. Springer, 2016, 885–892.
- Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J. and Tang, J. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *TACL* 9, 2021, 176–194.
- Xin, J., Lin, Y., Liu, Z. and Sun, M. Improving neural fine-grained entity typing with knowledge attention. In *Proceedings of AAAI'2018*, 5997–6004.
- Xiong, C., Power, R. and Callan, J. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of WWW'2017*, 1271–1279.
- Zeng, W. et al. PanGu-alpha: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation. 2021; arXiv:2104.12369.
- Zhang, Z., Han, X., Liu, Z., Xin Jiang, X., Sun, M. and Liu, Q. ERNIE: Enhanced language representation with informative entities. In *Proceedings of 2019 ACL*, 1441–1451.
- Zhang, Z. et al. CPM: A large-scale generative Chinese pre-trained language model. 2020; arXiv:2012.00413.

**Xu Han** is a Ph.D. candidate in the Department of Computer Science and Technology, Institute for Artificial Intelligence, and State Key Lab on Intelligent Technology and Systems at Tsinghua University, Beijing, China.

**Zhengyan Zhang** is a Ph.D. student in the Department of Computer Science and Technology, Institute for Artificial Intelligence, and State Key Lab on Intelligent Technology and Systems at Tsinghua University, Beijing, China.

**Zhiyuan Liu** is an associate professor in the Department of Computer Science and Technology, Institute for Artificial Intelligence, and State Key Lab on Intelligent Technology and Systems at Tsinghua University, Beijing, China.

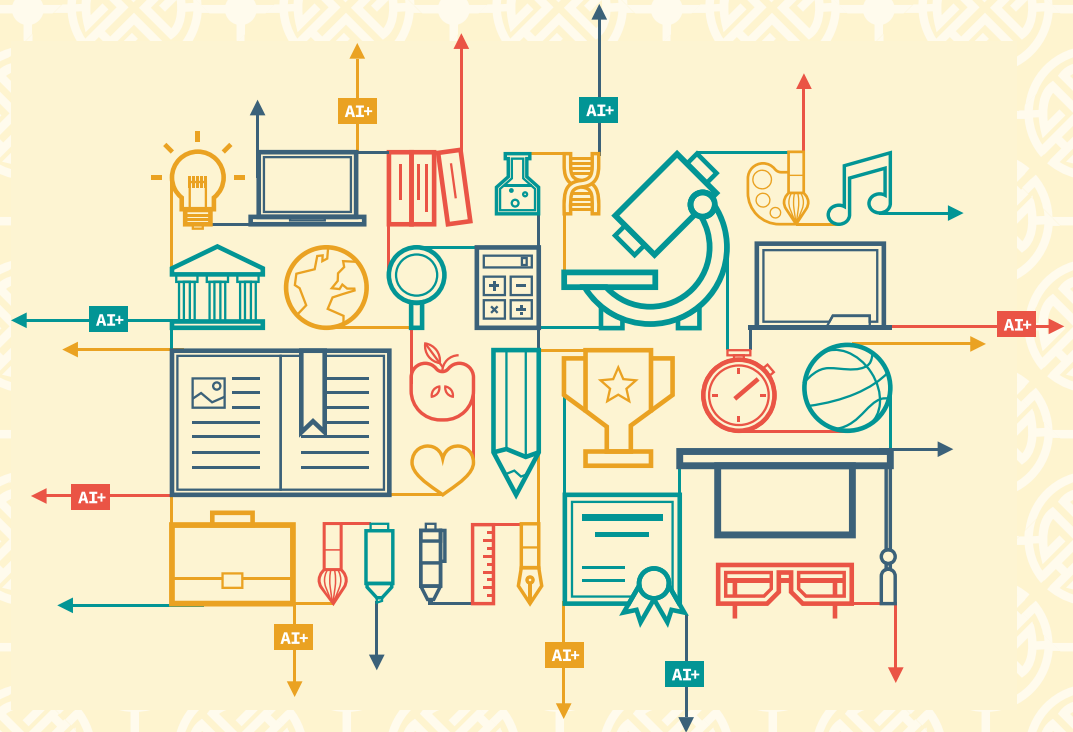
# AI+X Micro-Program Fosters Interdisciplinary Skills in China

BY FEI WU, QINMING HE, AND CHAO WU

**A**RTIFICIAL INTELLIGENCE (AI) has the potential to enhance every

technology as it resembles enabling technologies like the combustion engine or electricity. Many people in this field believe AI is general purpose, with a multitude of applications across many different disciplines. We believe the nature of AI is interdisciplinary. In other words, the power of AI lies in augmenting its ability to accelerate research exponentially and the possibilities are endless.

As a result, demand for professionals who are hard-wired in AI technology knowledge but who also possess interdisciplinary perspectives and transferable skills is becoming increasingly important. This article explores the endeavor of nurturing an



educational ecosystem to foster AI+X education in China via an interdisciplinary initiative.

AI is committed to the realization of machine-borne intelligence. The appropriate utilization of AI to a variety of re-

search fields is speeding up multiple digital revolutions, from shifting paradigms in health care, to education offered worldwide, to future cities made optimally efficient by autonomous vehicles. However, contemporary AI systems are good at specific predefined tasks and are unable to learn by themselves from data or from experience, intuitive reasoning, and adaptation. From the perspective of overcoming the limitations of existing AI, interdisciplinary scientific efforts are necessary to boost future research in this field. As a result, the next AI breakthroughs will be endeavors that draw upon neuroscience, physics, mathematics,

electronic engineering, biology, linguistics, and psychology to deliver great theoretical, technological, and applicable innovations, address complex societal issues, reshape the national industrial system, and more.<sup>2,6,8</sup>

## AI Undergraduate and Graduate Curricula

AI has become an undergraduate major at more than 300 universities in China as of March 2021, as approved by the Ministry of Education in 2019. Many Chinese universities established AI schools (such as Xidian, Nanjing University, and Xi'an Jiaotong) and institutes (such as Zhejiang University, Peking University, and Tsinghua University) in

**China is fostering AI education in universities by strengthening the interdisciplinary links between AI and relevant fields, instead of merely offering a few core courses as a part of the CS discipline.**

recent years for research training, especially AI Ph.D. programs. Today, AI is the fastest-growing discipline at China's universities. China is fostering AI education in universities by strengthening the interdisciplinary links between AI and relevant fields, instead of offering a few core courses as a part of the computer science discipline. For example, China's top music university—Central Conservatory of Music—started to recruit students with the three-year Ph.D. program covering music and AI, and Southwest University of Political Science & Law opened a school of AI and law to equipped students with strong AI+ law knowledge and professional skills.

Since the fundamental goals of AI are to enable machines with human-like capabilities, such as sensing (for example, speech recognition, natural language understanding, computer vision), problem-solving (for example, search, optimization, and learning) and acting (for example, robotics and systems), the core AI courses in undergraduate and graduate curricula generally consist of computer science, mathematics, and statistics. In particular, AI ethics and responsibility to humankind are an important part of the courses since the long-term goal of keeping AI beneficial to human society is crucial.

### AI+X Micro-Program

A micro-program consists of a set of micro-courses in a specified field. Micro-courses can help learners gain knowledge in small units in a short period of time. As a result, micro-

courses provide potential opportunities for professionals to expand their competencies without leaving their current roles and are designed to keep their expertise up to date.

In April 2021, an AI+X micro-program was offered to 300 students outside the AI discipline by six top universities in east China—Shanghai Jiao Tong University, Fudan University and Tongji University in Shanghai, Zhejiang University in Zhejiang Province, Nanjing University in Jiangsu Province, and the University of Science and Technology of China in Anhui Province, and with some companies such as Huawei, Baidu, and SenseTime. Since an AI+X micro-program is expected to train students with a solid background in a thematic discipline (X) who need AI to tackle a specific scientific challenge, the AI+X micro-program currently allows the students from thematic discipline to register.

The accompanying table lists the curricula of the AI+X micro-program, which consists of more than 40 online courses in six categories (prerequisite course, AI fundamental compulsory course, module course, algorithmic practical course, interdisciplinary course, and summer camp course). Each course is taught in online SPOC (small private online course) to registered students for 11 weeks. Each course is scheduled to meet 1–2 hours a week. After finishing each course, an offline examination is given.

Each registered student is required to complete a total of at least 12 credits within 1–2 years to obtain

a certificate signed together by the aforementioned universities. Moreover, these universities allow registered students to use AI+X online courses to count toward their degrees. That is, registered students who enroll in an AI+X micro-program may have those credits transferred to fulfill their major requirements.

Motivated by the mission of quality-first, with open and cooperative sharing, AI+X micro-program courses are expected to be available to the public in the near future.

### AI+X Education Ecosystem

The foundation for a healthy AI educational ecosystem is built on AI+X skills training, textbooks, online courses, and practice platforms as well as the collaboration between

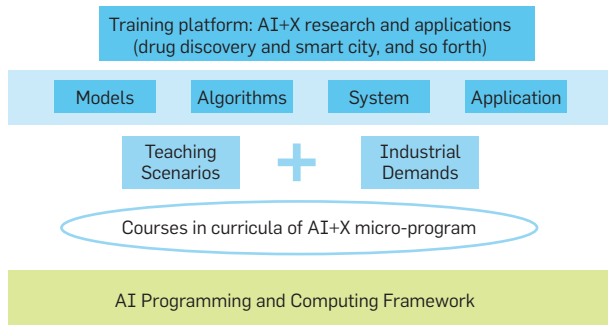
universities, government, and industry.

Textbooks are essential for students as well as teachers. For the student, an effective textbook offers guidelines for what they can expect to learn. For the teacher, it provides objective reasons for why a particular topic should be taught and the educational goals it satisfies. In cooperation with more than 30 universities, China's higher education press is publishing a series of AI textbooks that cover theoretical models, algorithms, technologies, AI ethics as well as AI+X interdisciplinary research.

Online courses allow students to learn from anywhere and at any time, which is more flexible, customized, and cost-effective than traditional education. Each course in the AI+X micro-program

#### The curricula of AI+X micro-program.

Category	Courses (in total more than 40 online courses)	Description
Prerequisite Course	Data structure, C programming language, problem-solving, Python, Java	No credit required. We encourage students from X disciplines to have basic programming skill before attending AI+X micro program
AI Fundamental Compulsory Course	The introduction to AI, pattern recognition and machine learning, AI programming, seminar for AI advanced topic	Each student is required to select three courses
Module Course	Currently there are six modules: sensing and perception, hardware and system, creative design, game and decision, robotics, and intelligent city,	Each module comprises several courses. Each student is required to select at least two courses from two modules.
Algorithmic Practical Course	Courses provided mainly by companies such as Huawei, Baidu, and SenseTime, among others.	Each student is required to select one course.
Interdisciplinary Course	Interdisciplinary course such as AI + health, AI + economics, AI + law, AI + pharmacy, AI + finance, AI + public management	Each student is required to select one course
Summer Camp Course	No credit required	



### The architecture of Wise Ocean.

is required to be available online and some courses have been recognized as national-level, high-quality MOOCs.

The best way to learn AI is to practice. One cannot truly learn until and unless one truly gets some hands-on training for solving real problems. The training platform used in the AI+X micro-program is *Wise Ocean*, which provides a one-stop repository of resources to help AI+X curriculum learners better understand the promise and implications of domain-specific AI, such as machine-learning driven drug discovery (see the accompanying figure).

Wise Ocean, like MINIX, is not merely a teaching tool. Early versions of MINIX were created merely for educa-

tional purposes. Starting with MINIX 3, the primary aim of development shifted from education to the creation of a highly reliable and self-healing microkernel OS. The Wise Ocean platform would collect open source codes for plenty of AI+X interdisciplinary innovative research since registered students will finish their assigned projects. Therefore, Wise Ocean will evolve into a synergistic combination of research and education—a platform that can educate students (future innovators) to apply cutting-edge AI technologies to many industries.

In March 2020, China's Ministry of Education, the National Development and Reform Commission, and the Ministry of Finance co-issued guide-

**Wise Ocean will evolve into a synergistic combination of research and education—a platform that can educate students (future innovators) to apply cutting-edge AI technologies to many industries.**

lines to urge universities to create interdisciplinary skill-cultivation systems to substantially improve the level of graduate education in AI and also encourage enterprises to ramp up investment to support the development of AI-related disciplines and high-level skills training. The AI+X micro-program sets up a collaborative innovation system via interactions with universities and industry for AI development. For example, companies provide their programming tools, industrial demands, and computing power to cultivate AI+X students.

### Challenges and Conclusion

Most research fields increasingly encompass data analysis that requires computer science skills. The University of Illinois has designed a set of CS+X degree programs that allow students to pursue a flexible program of study incorporating a strong grounding in computer science with technical or professional training in the arts and sciences (such as CS+ Astronomy, CS+ Chemistry, and CS+ Economics). AI+X micro-programs in China must practice online (MOOC or SPOC), resulting in several challenges:

1. Qualified textbooks as well as the corresponding online courses to guide students to better learn online courses.
2. A flexible training platform consisting of data, model, teaching scenario, and industrial demands.
3. Computing power for students to train their AI models online is difficult, but rising faster than ever before.

4. Bridging the gap between AI+X interdisciplinary research to offer suitable AI+X teaching projects for students not from AI discipline.<sup>1,3,4</sup>

Accelerating the development of a new generation of AI is a key strategy for China, to boost developments in science and technology, to upgrade every industrial domain, and to increase overall productivity.<sup>5,7</sup> Building up an AI ecosystem is very important to nurture more AI+X talents. In a healthy AI ecosystem, each participant across multiple industries and domains can find various ways to thrive.

### References

1. Fan, J., Fang, L., Wu, J., Guo, Y., Dai, Q. From Brain Science to Artificial Intelligence. *Engineering* 6, 3, (2020), 248–252.
2. Lyu, Y.-G. Artificial intelligence: Enabling technology to empower society. *Engineering* 6, 3 (2020), 205–206.
3. Pan, Y. Multiple knowledge representation of artificial intelligence. *Engineering* 6, 3, (2020), 216–217.
4. Pan, Y. Miniaturized five fundamental issues about visual knowledge. *Front Inform Technol Electron Eng* 22, 5, (2021), 615–618.
5. Wu, F. et al. Towards a new generation of artificial intelligence in China. *Nature Machine Intelligence* 2 (2020), 312–316.
6. Wu, W., Huang, T., Gong, K. Ethical principles and governance technology development of AI in China. *Engineering* 6, 3 (2020), 302–309.
7. Zhu, J., Huang, T., Chen, W., Gao, W. The future of artificial intelligence in China. *Commun. ACM* 61, 11 (Nov. 2018), 44–45.
8. Zhuang, Y., Cai, M., Li, X., Luo, X., Yang, Q. and Wu, F. The next breakthroughs of artificial intelligence: The interdisciplinary nature of AI. *Engineering* 6, 3, (2020), 245–247.

**Fei Wu** is a professor in the College of Computer Science at Zhejiang University, Hangzhou, China.

**Qinming He** is a professor in the College of Computer Science at Zhejiang University, Hangzhou, China.

**Chao Wu** is an associate professor in the School of Public Affairs at Zhejiang University, Hangzhou, China.

This work is supported by the NSFC (61625107, 62037001).

© 2021 ACM 0001-0782/21/11



# AI Start-Ups in China

BY JING YANG

**C**HINESE AI BUSINESSES have been growing rapidly since 2010. They have attracted significant investment from Internet giants and a vast number of emerging AI companies have emerged. Over the past decade, Chinese AI start-ups have gradually moved away from noisy bubbles and landed in an investment boom. In 2020, when people were fighting against the pandemic, CloudMinds, an AI start-up based in Beijing, developed a humanoid service robot named Cloud Ginger XR-1. Ginger played an important role in local hospitals, delivering food and medication to patients in a contactless manner when it was needed the most. Moreover, Ginger entertained patients, freeing up doctors and medical teams to focus on more critical health matters.

**China ranked second worldwide in total investments in AI start-ups.** China's AI start-up

ecosystem was lagging behind U.S. in 2017, according to a McKinsey Global Institute analysis of the country's AI sector.<sup>1</sup> However, the mountains of data available due to China's population served as a precondition for training AI systems, so the country's global ranking soon soared.

The AI industry took a hit with the COVID-19 outbreak, but statistics show it is still on a growth trend. In 2020, China also invested \$9.9 billion in AI start-ups, accounting for 23.52% of global investment and ranking second.<sup>2</sup>

China saw a very substantial increase in start-up investments in 2017 and 2018, reaching a maximum of \$12.6 billion in 2018. However, it declined in 2019 and 2020 due to the impact of the pandemic.<sup>2</sup>

**Chinese AI start-ups have a relatively concentrated distribution.** By the end of 2019, there were more than 1,000 active AI companies in mainland China, accounting for



Cloud Ginger

more than one-fifth the global number. Beijing-owned AI businesses (468) ranked first globally. Among the top 20 cities in the world with the most significant number of AI companies, there are four located in China. The AI companies in China are primarily based in the Beijing-Tianjin area, Yangtze River Delta, Pearl River Delta, and the Midwest's key provinces. There are several fields, including computer vision, intel-

ligent speech and semantics, robotics, and drones, that have shown a tendency toward industrial clusters.

The Chinese AI industry targets three main layers: Basics, technology, and application. Some 80% of Chinese AI start-ups focus on the application layer, and a great percentage of those companies cover robotics, drones, AI and healthcare, AI and finance, and AI and manufacturing. Moreover, there are

**Over the past decade, Chinese AI start-ups have gradually moved away from noisy bubbles and landed in an investment boom.**

# Publish Your Work Open Access With ACM!

ACM offers a variety of Open Access publishing options to ensure that your work is disseminated to the widest possible readership of computer scientists around the world.



Please visit ACM's website to learn more about ACM's innovative approach to Open Access at: [www.acm.org/openaccess](http://www.acm.org/openaccess)



Association for  
Computing Machinery

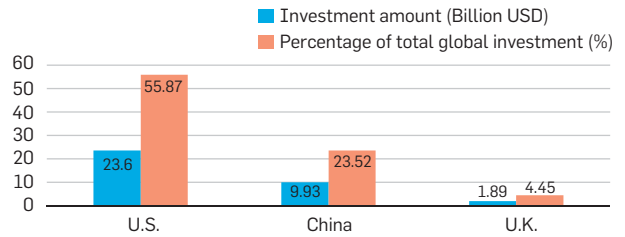


Figure 1. Comparison of the amount of investment in AI start-ups in the U.S., China, and the U.K. in 2020.

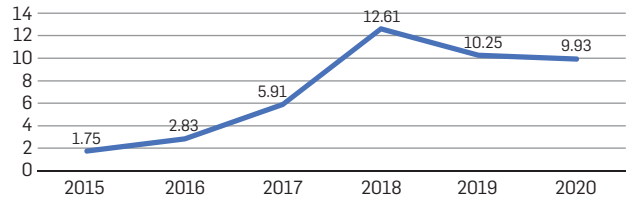


Figure 2. China's investment amount in AI start-ups from 2015–2020.

unicorn companies, for example, Megvii, SenseTime, DJI, and UBTECH, that also fall under the applications umbrella.

**China reshapes the industry and blossoms in chip R&D and autonomous vehicle start-ups.** As predicted, the semiconductor industry experienced hardships in 2021 as the chip shortage escalated. China has attached great importance to developing an independent industrial chain with core technology and regards it as an effective national strategy. Indeed, original innovation has become the new engine of AI entrepreneurship in China. Start-ups including Cambricon, Yuntian Lifei, and Horizon have scaled in AI chips. Biren and Suiyuan have received major funding. Companies like Yitu, which started with algorithms, have also stepped into the chip industry.

Autonomous driving is another area where China's AI start-ups have exploded. Didi, for example, is an established autonomous vehicle

company. NIO, Xiaopeng, and Li Auto are accelerating the pace of developing new energy vehicles. Moreover, Deepblue AI, Pony AI, and other companies building self-driving cars have started a new funding round. With this massive injection of capital, China's AI industry has seen a large number of unicorns as well as industrial agglomeration in the fields of computer vision, semiconductor, and autonomous vehicles. It is believed that in the near future, AI enterprises will blossom across the entire continent and AI entrepreneurship will show a large-scale outbreak with more high-tech companies emerging! 

#### References

1. Barton, D., Woetzel, J., Seong, J. and Tian, Q. *Artificial Intelligence: Implications for China*. McKinsey Global Institute, 2017.
2. Zhang, D. et al. *Artificial Intelligence Index Report 2021*. AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, USA, 2021.

Jing Yang is the founder and CEO of AI Era, Beijing, China.

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

# Natural Interactive Techniques for the Detection and Assessment of Neurological Diseases

BY FENG TIAN, YUNTAO WANG, AND YICHENG ZHU

**N**EUROLOGICAL DISEASES, SUCH as cerebrovascular disease, Parkinson's disease (PD), Alzheimer's disease, have become the leading cause of death in China. Neurological function evaluation is crucial for the diagnosis and intervention of neurological diseases. Clinically, neurological function is evaluated by various scales, tests, and questionnaires. However, these methods rely on costly professional equipment and medical personnel. They cannot be used as a means of daily evaluation of neurological diseases. Natural user interface (NUI) is a new generation of user interfaces. In recent years, NUI-based health sensing has become a hot topic in human-computer interaction research. This article discusses recent advances in the use of NUI for neurological disease detection and assessment. We also share some of our practices in this area.

NUI enables strong perception, natural information access, and multichannel ability of computing systems, which can open new opportunities on quantitative, multimodal, and implicit monitoring support for detecting neurological diseases. Ubiquitous computing and multimodal sensing are two NUI technologies that provide NUI with distinct advantages in disease detection over conventional neurological assessment approaches.

Ubiquitous computing allows us to have access to robust sensing of human physiological states anywhere and anytime. Enabled with cutting-edge machine learning and signal processing approaches, the ubiquitous device can sense the "hidden" physiological signal using its built-in sensors. Therefore, we can provide real-time physiological signals as inputs to communicate with the user interface, forming a bio-

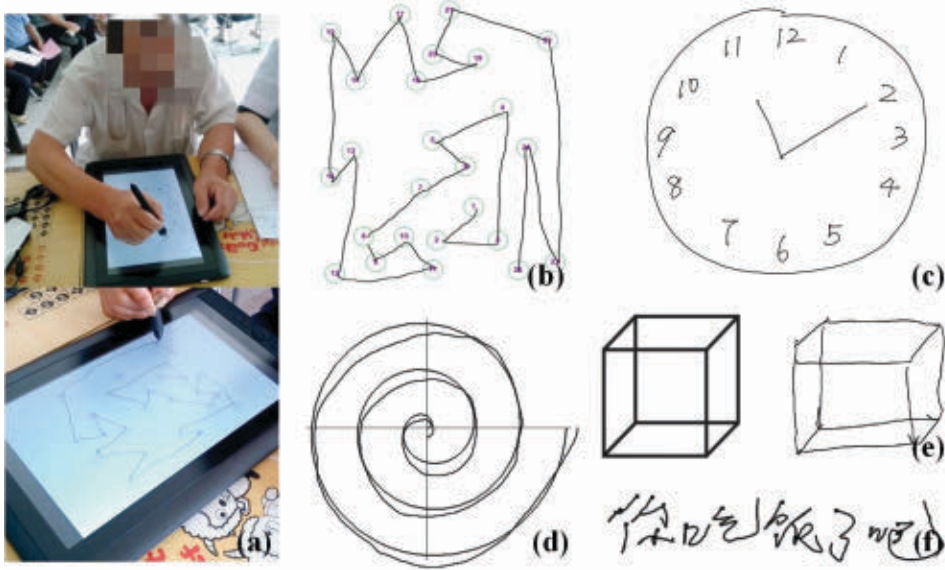


cybernetic loop between the user and the computing system. Harvard Sensor Lab developed a Mercury system that consists of several wearable accelerometers for long-term data collection for people affected by neurological diseases.<sup>6</sup> The ability to acquire, process, and wirelessly transmit physiological data during daily activities is the primary advantage of ubiquitous computing technologies.

The multimodal sensing ability in NUIs also provides unique power in detecting neurological diseases. Symptoms of such diseases vary among people. Some

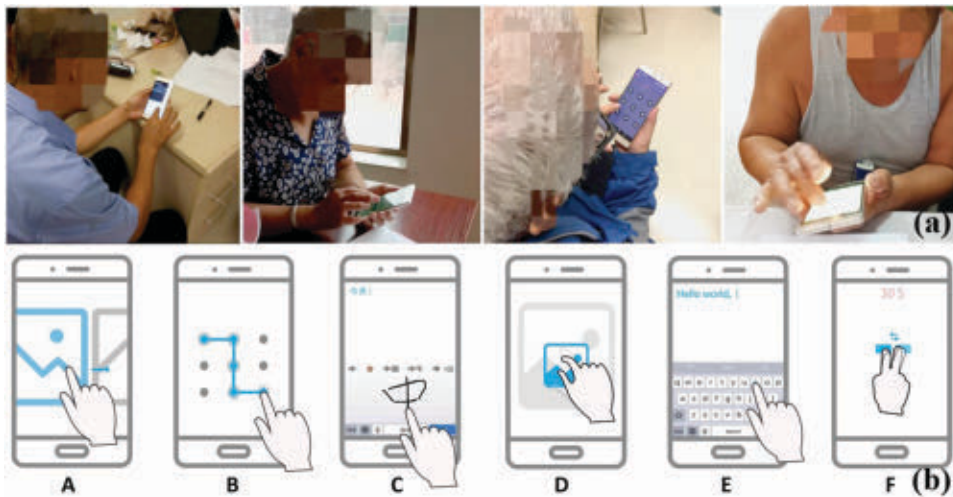
patients with neurological diseases have motor deficits in gait or hands, while others show cognitive impairments such as memory loss or difficulty in decision-making. NUI-based diagnosis systems use multimodal input such as voice, finger touch, body motion, and facial expression to evaluate neurological function. By combining the advantages of all these inputs in sensing different symptoms, the NUI-based diagnosis systems can provide more comprehensive and accurate assessments of patients' neurological function. Researchers found

**The multimodal sensing ability in NUIs also provides unique power in detecting neurological diseases.**



a) Subject is taking the test; b) trail-making test; c) clock drawing test; d) Archimedes spiral test; e) Graph Copying Test; f) handwriting test.

**Figure 1. Pen-based nervous system disorders system and typical test.**



a) Sample pictures of subjects in this study; b) Common touch gestures explored in this research. A: flick gestures; B: drag gestures; C: handwriting gestures; D: pinch gestures; E: tap gestures (typing); F: alternating finger tapping (AFT, included as a clinical reference).<sup>9</sup>

**Figure 2. Detecting motor impairment on smartphones.**

relationships between patients' performances and neurological function in many interaction modalities. Speech pathologists found there is a relationship between neurological function and pronunciation.<sup>7,12</sup> Sketching can reflect motor and cognitive function impairments in patients with neurological

diseases.<sup>2,8,10</sup> Gait, as a behavioral feature of human's daily movement, can reflect human mobility, physical fitness, and health,<sup>3</sup> and can thus be used in PD detection.<sup>5</sup> A variety of approaches based on smartphones have been proposed to detect nervous system diseases. These approaches use embedded

IMU<sup>1</sup> or typing activities<sup>4,11</sup> in smartphones to detect motion abnormalities of neurological diseases. The multimodal sensing ability of NUI provides key technical support for the whole process of diagnosis and treatment, such as early warning screening, clinical diagnosis, prognosis evaluation, rehabilitation

monitoring, and long-term tracking for neurological function evaluation.

Our joint research group from the Institute of Software, Chinese Academy of Sciences, and Peking Union College Hospital have developed a set of multi-modal natural human-computer interactive diagnosis tools for neurological disease detection. We use pens, postures, intelligent objects, voice, touch-screen mobile devices, and other multichannel interactive devices to carry out early warnings and auxiliary diagnoses of nervous system diseases. Therefore, we can enable health screening, clinical diagnosis, prognosis evaluation, and rehabilitation monitoring for neural health.


We explored diagnosing central nervous system disorders from a wide range of pen-based sketching tasks (see Figure 1). We proposed novel approaches to extract features that reflect both motion functions and cognitive functions of the human body and are independent of the content and type of the sketches. In a 490-subject (107 patients) user study, we found our approach achieved 83.15% average accuracy on five sketching tasks with different degrees of freedom. The result demonstrated the feasibility of diagnosing neurological diseases via daily sketching activities.

We also explored the feasibility and accuracy of detecting motor impairment in early PD via sensing and analyzing users' common touch gestural interactions on smartphones (see Figure 2). We investigate four common gestures, including flick, drag, pinch, and handwrit-

## NUI-based diagnosis systems use multimodal input such as voice, finger touch, body motion, and facial expression to evaluate neurological function.

ing gestures, and propose a set of features to capture PD motor signs. Through a 102-subject (35 early PD subjects and 67 age-matched controls) study, our approach achieved an AUC of 0.95 and 0.89/0.88 sensitivity/specificity in discriminating early PD subjects from healthy controls.<sup>9</sup> Our work constitutes an important step toward unobtrusive, implicit, and convenient early PD detection from routine smartphone interactions.

This project was selected as one of the “30 best cases of AI applications in the medical and health domain” by the national health department. The previous achievements in this project won the second prize in national science and technology progress in 2018.

The existing technical achievements and our practice show that the strong perception, naturalness information access, and multi-channel ability of NUI can facilitate neurological assessment methods and provide quantitative, multimodal, and non-task monitoring support to detect neurological diseases. 

### References

1. Carignan, B., Daneault, J.-F., and Duval, C. Measuring tremor with a smartphone. *Mobile Health Technologies*. Springer, 2015, 359–374.
2. Davis, R. et al. THink: Inferring cognitive status from subtle behaviors.

3. Drake, J.M., Griffen, B.D. Early warning signals of extinction in deteriorating environments. *Nature* 467, 7314 (2010), 456–459
4. Iakovakis, D., Hadjidimitriou, S., Charisis, V., Bostantzopoulou, S., Katsarou, Z., and Hadjileontiadis, H.J. Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage Parkinson's disease. *Scientific Reports* 8, 1 (2018), 7663.
5. Lan, K.-C., Shih, W.-Y. Early diagnosis of Parkinson's disease using a smartphone. *Procedia Computer Science* 34 (2014) 305–312.
6. Lorincz, K. et al. Mercury: A wearable sensor network platform for high-fidelity motion analysis. *SensSys* 9 (2009) 183–196
7. Roy, N. et al. Evidence-based clinical voice assessment: a systematic review. *American J. Speech-Language Pathology* 22, 2 (2013), 212–226.
8. Smits, E.J. et al. Standardized handwriting to assess bradykinesia, micrographia and tremor in Parkinson's disease. *PLOS ONE* 9, 5 (2014), e97614
9. Tian, F. et al. What can gestures tell? Detecting motor impairment in early Parkinson's from common touch gestural interactions. In *Proceedings of the 2019 CHI Conf. Human Factors in Computing Systems*, 1–14.
10. Ünlü, A., Brause, R., Krakow, K. Handwriting analysis for diagnosis and prognosis of Parkinson's disease. In *Proceeding of Intern. Symposium Biological and Medical Data Analysis*. Springer Berlin Heidelberg, 2006, 441–450.
11. Wang, Y., Yu, A., Yi, X., Zhang, Y., Chatterjee, I., Patel, S., Shi, Y. Facilitating text entry on smartphones with QWERTY keyboard for users with Parkinson's disease. In *Proceedings of the 2021 CHI Conf. Human Factors in Computing Systems*, 1–11.
12. Whiting, S., Rydell, R., Åhländer, V.L. Design of a clinical vocal loading test with long-time measurement of voice. *J. Voice* 29, 2 (2015), 13–261.

**Feng Tian** is a professor at the State Key Laboratory of Computer Science and Beijing Key Lab of Human-Computer Interaction in the Institute of Software at Chinese Academy of Sciences, Beijing, China.

**Yuntao Wang** is an assistant professor at the Department of Computer Science and Technology, Tsinghua University in Beijing, China.

**Yicheng Zhu** is a professor at Peking Union Medical College Hospital in Beijing, China.

© 2021 ACM 0001-0782/21/11

## ACM Transactions on Computing for Healthcare (HEALTH)

*A multi-disciplinary journal for high-quality original work on how computing is improving healthcare*

ACM Transactions on Computing for Healthcare (HEALTH) is the premier journal for the publication of high-quality original research papers, survey papers, and challenge papers that have scientific and technological results pertaining to how computing is improving healthcare.



For further information and to submit your manuscript, visit [health.acm.org](http://health.acm.org)



# Processing Extreme-Scale Graphs on China's Supercomputers

BY YIMING ZHANG, KAI LU, AND WENGUANG CHEN

**M**ANY APPLICATIONS, SUCH as Web searching, social network analysis, and power grid management, require extreme-scale graph processing. However, it is very challenging because graph processing exhibits unique characteristics such as more load imbalance, lack of locality, and access irregularity. The extreme graph scale makes the situation even worse.

Supercomputers have large compute, large memory, and fast interconnect. They seem ideal for extreme-scale graph processing. China has built several leading supercomputers in the world. For example, the Tianhe-2 and Sunway TaihuLight supercomputers have ranked No. 1 in Top500 list 10 times between June 2013 through May 2018. Like many other supercomputers, they use heterogeneous accelerators to achieve high performance

**One of the unique features of the TaihuLight machine is the heterogeneous on-chip SW26010 CPU.**



The TaihuLight supercomputer at the National Supercomputer Center in China's Jiangsu province.

for regular workloads such as stencil-based and structured grid-based computations. Are they also capable of processing extreme-scale graphs?

In this article, we discuss our efforts to enable extreme-scale graph processing in two leading supercomputer architectures.

**Tianhe hardware features.** The Tianhe supercomputer has unique designs for its many-core

CPU architecture and its high-performance interconnection network, providing both opportunities and challenges for graph processing.

Tianhe's computing nodes (CN) use the proprietary Matrix-2000+ CPUs (see Figure 1). A CN has three Matrix-2000+ CPUs, each of which has 128 2GHz cores. Each core has an in-order 8-to-12-stage pipeline extended with scalable vector extension (SVE). Matrix-2000+ CPUs adopt a regional autonomous parallel architecture where one CPU is composed of four regions connected through a scalable on-chip communication network. Each region is a functionally independent supernode (SN) with four panels communicating through an intra-region

interface. Each panel has eight cache-coherent compute cores. SVE enables Matrix-2000+ to choose the most appropriate vector length via two usage modes: auto vector-length agnostic (AVLA) mode and assembly vector-length specified (AVLS) mode. AVLA mode can automatically pack sub-vectors into vectors but requires synchronization between processing of two vectors, while AVLS mode allows programmers to specify user-defined sub-vector lengths.

Tianhe's network subsystem adopts a multi-dimensional tree topology with optoelectronic-hybrid interconnection, which combines the benefits of both tree and n-D-Torus topologies. The networking logic is integrated into

the network interface chip (HFI-E) and the network router chip (HFR-E). HFI-E implements the proprietary MP/RDMA (mini packet/remote direct memory access) communication and collective offloading mechanism. CNs connected to an HFR-E are in the same communication domain. Tianhe has highly optimized its intra-domain communication, which is an order of magnitude faster than its inter-domain communication crossing multiple HFR-Es.

**Leveraging hardware features for graph processing.** Different from the traditional SIMD (single-instruction-multiple-data) technique, the Matrix2000+ CPUs support vectorization to accelerate graph computation. We leverage SVE to realize efficient graph traversal.

Traditional vectorization induces synchronization between processing consecutive vectors (by inserting stalls) and thus lowers the overall performance. Fortunately, we find that graph traversal (such as BFS) simply scans a vertex range to determine the vertices to-be-traversed at the next level, allowing avoiding synchronization if none of the vertices belongs to more than one level. This situation might

exist only because of the existence of loops in the graph. To address the loop problem, if a vertex exists in two successive levels and causes a loop, we split it into two virtual ones. Moreover, if we find a vertex that belongs to multiple levels during pre-processing, we use a virtual vertex at each level to participate in that level's vectorized processing.

We adopt AVLA and AVLS to realize graph traversal efficiently. If the traversal is likely to encounter loops (for example, in top-down BFS), then we adopt AVLA to automatically pack unvisited neighbor vertices (sub-vectors) into vectors while avoiding vertex splitting (at the cost of explicit synchronization). Otherwise, it is unlikely to encounter loops (for example, in bottom-up BFS), and thus we pack the neighbors into the vectors through AVLS and split vertices once loops occur. AVLA and AVLS accelerate the procedure that every thread (core) handles a different vertex range and examines the edges connected to unvisited vertices, so as to determine whether the neighbor vertices should be visited on the next level.

To adapt graph processing to the topology

## The Tianhe supercomputer has unique designs for its many-core CPU architecture and its high-performance interconnection network, providing both opportunities and challenges for graph processing.

of Tianhe's multi-dimensional tree network, we refactorize the graph with *fusion* and *fission*<sup>3</sup> when storing graph vertices and edges. Specifically, fusion organizes a set of neighboring low-degree vertices into a super-vertex (for processing locality), and fission splits a high-degree vertex into a set of sibling sub-vertices (for load balancing). Refactorization is performed in parallel by all workers on CNs, and we resolve the potential conflict by double-checking on vertex states. A worker checks whether the vertex has been processed both before moving it to the processing queue and before performing fusion/fission. If a conflict occurs, further processing will be skipped.

The vertices and edges of the refactorized graphs are assigned to the CNs in the network according to the proximity of the multi-dimensional tree topology. Neighboring vertices are assigned to the same communication domain. Graph refactorization could be pipelined with assignment and thus only induces a small extra overhead.<sup>3</sup> Note the partitioning results are reusable. Thus, it is worth paying for

the extra refactorization overhead in most cases.

We further leverage the topology information to perform aggressive message aggregation. Messages are gathered to the responsible nodes (referred to as monitors) in the source domains, transferred between monitors, and scattered to the target nodes in the target domains. We adopt adaptive buffer switching and dynamic buffer expansion to reduce communication cost effectively.

**Applications of graph processing on Tianhe.** The Tianhe graph processing system has been widely used in industry throughout China in areas such as computational biology, industrial simulation, and visualization.

Beijing Genomics Institute (BGI) and Shanghai Institute of Materia Medica (SIMM) jointly constructed a high-throughput drug virtual screening platform based on Tianhe graph processing system. The platform screened over 40 million molecular compounds for anti-Ebola virus drugs per day, achieves the fastest high-throughput virtual drug screening in history, and plays an im-

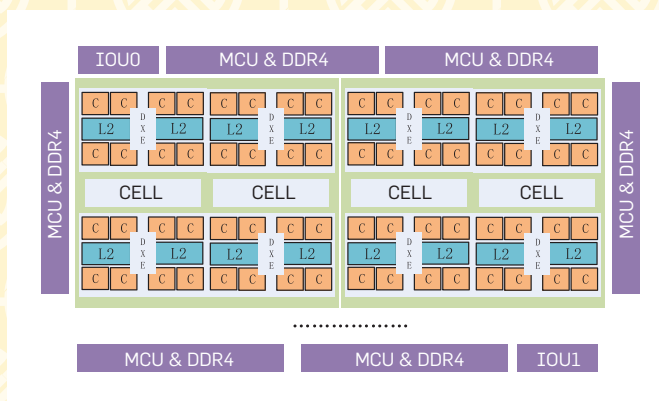


Figure 1. Matrix-2000+ CPU architecture.

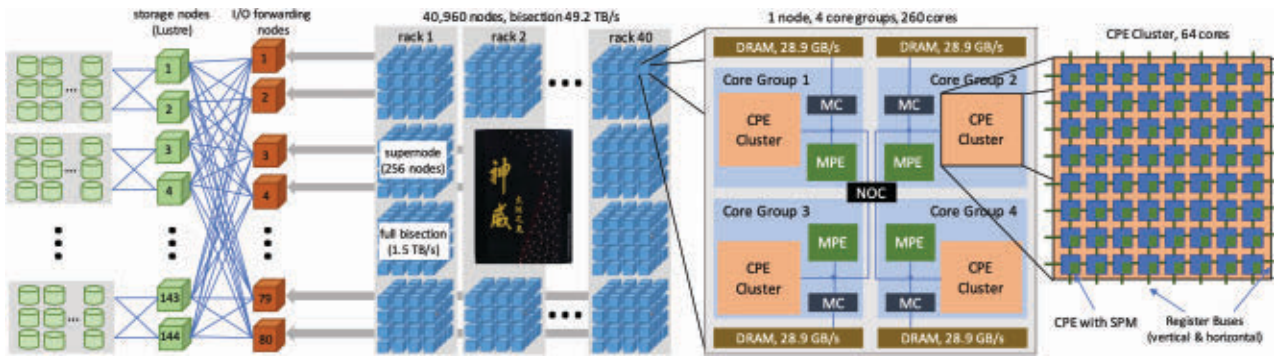


Figure 2. The architecture of TaihuLight.

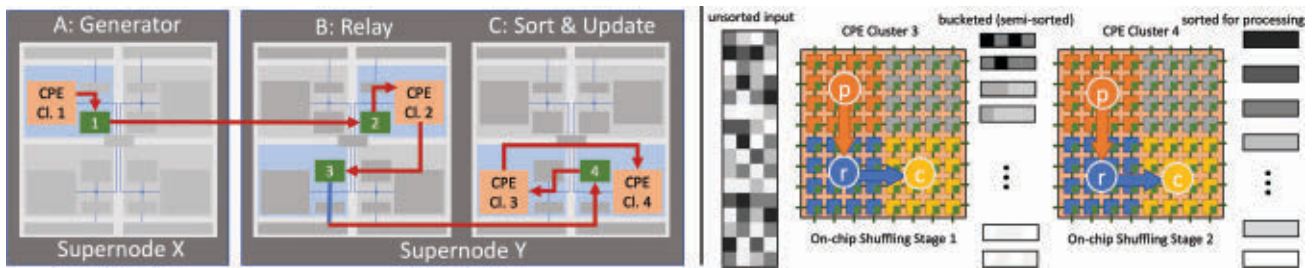


Figure 3. Supernode routing and module mapping in CPEs.

portant role in anti-Ebola drug development. The core algorithm, Lamarckian Genetic Algorithm (LGA),<sup>2</sup> is formed as a global-local-hybrid search problem and deeply accelerated on Tianhe graph processing system using 512 ~ 8192 CNs (up to 19.7K CPU cores). The efficiency of our graph processing system reaches as high as 60%.

Shaanxi Key Laboratory of Large-scale Electromagnetic Computing

(LEC) has developed the complex matrix local block pivoting LU (LBPLU) decomposition software for applications of massive parallel Method of Moments (MoM). The LEC laboratory ran LBPLU with local pivoting on Tianhe to solve the matrix equation generated by MoM. Specifically, for simulating the electromagnetic scattering of an aircraft, LBPLU divides the aircraft surface into a grid structure, where each grid node

is a vertex of a graph, and the influence between grid nodes is modeled as edges. LBPLU performs triangulation on the grid to realize flow visualization. When running LBPLU, our system achieves as high as 50.16% parallel efficiency when the parallel scale is 8192CNs (19.7K cores).

We have also deployed the graph processing system on a subset of the next-generation exascale Tianhe supercomputer, which consists of 512CNs with 96608 cores. Performance evaluation shows that the 512-node sub-system achieves 2131.98 Giga TEPS (traversed edges per second) for the BFS test of Graph500, which outperforms Tianhe-2 Supercomputer with 16x more nodes.

in Figure 2. One of the unique features of TaihuLight machine is the heterogeneous on-chip SW26010 CPU (see right part of Figure 2). Each SW26010 CPU comprises four core groups (CGs) connected via a low-latency on-chip network (NoC). Each CG consists of a management processing element (MPE), a 64-core computing processing element (CPE) cluster, and a memory controller (MC), and thus a total of 260 cores per CPU (node). Each CPE comes with a 64KB scratch pad memory (SPM) without cache, which requires explicit programmer control. The architecture demands manual coordination of all data movement, which is a particularly challenging task for irregular random accesses.

The TaihuLight CNs are connected via a 2-level In-

**To maximize utilization of the full-bisection intra-supernode bandwidth, we form target groups using supernode boundaries.**

**Processing Graphs on Sunway TaihuLight**

The architecture of TaihuLight is illustrated



finiBand network. A single-switch with full bisection bandwidth connects all 256 nodes within a supernode, while a fat-tree with 1/4 of the full bisection bandwidth connects all supernodes (see left part of Figure 2).

### Mapping the graph processing modules to heterogeneous processors.

Within each SW26010 CPU, the four CGs are assigned with distinct functions as shown in Figure 3: (A) Generation, (B) Relay, (C1) Coarse sort, and (C2) Update. This function mapping is static, and each function is performed by one CG only. The goal of this mapping is to achieve balanced CG utilization. This pipelined architecture allows us to process batched data in a streaming way, gaining lower I/O complexity to main memory and higher utilization of the on-chip bandwidth.

At the second level of specialization, we leverage the specific hardware features within each CG. The MPE is well suited for task management, plus network and disk I/O, while the CPEs are tightly connected through the 2D fast communication feature, naturally leading us

to assign communication tasks on the MPE and data sorting tasks on the CPEs.

**Supernode routing.** This technique targets efficient inter-node communication to enable our heterogeneous processing pipeline on the full system.

The performance of distributed graph applications are usually damaged by large numbers of small messages sent following the graph topology. The all-to-all style small messages among 10,000s of nodes is inefficient due to per-message overheads (routing information, connection state, and so on.) We propose a supernode routing technique to mitigate this by factoring all compute nodes into groups according to their supernode affiliation. Each node combines all messages to nodes within the same target group into a single message sent to a designated node within that group. This so-called relay node unpacks the received messages, combing messages from different source groups, repack the messages to each in-group target node into one message, and distributes them to appropriate peers.

To maximize utilization

of the full-bisection intra-supernode bandwidth, we form target groups using supernode boundaries. Each source node minimizes the number of relay nodes it sends to within a target group (usually one relay node per target group) to perform message aggregation effectively. To achieve load balance, each node in a target group acts as a relay node. The situation is more complicated if there are failure nodes in a supernode, and we use a stochastic replay assignment to maintain load balance.


**The Shentu graph processing framework.** In addition to the hardware specialization and supernode routing, we also have other innovations such as on-chip CPU sorting and degree-aware messaging. We omit them here due to limited space. Readers interested in more details should refer to Lin et al.<sup>1</sup>

Finally, we designed and implemented a vertex-centric graph processing framework, Shentu, in Sunway TaighuLight. It could support graph algorithms such as PageRank, WCC, and BFS with around 30 lines of code to run on the full system of TaighuLight.

It should be noted that the Sogou graph is

the largest real graph processed in literature, which has 12 trillion edges and is prohibitive for small scale systems. Shentu could process it with 8.5s for each iteration of PageRank in full scale Sunway TaighuLight (see the accompanying table). The 42.kron (70 trillion edges) is also the largest synthetic graph processed in literature to date.

### Conclusion

In this article, we showed how graph processing is efficiently supported by supercomputers with different heterogeneous architecture characteristics: on-chip processing element array with SPMs and wide vector units. We also showed how techniques such as vectorization and supernode routing are used to optimize the all-to-all messages of graph computing. We expect the result not only enables extreme-scale graph processing, but also hints at the possible fusion of supercomputing and big data architectures. 

### References

1. Lin, H., et al. ShenTu: Processing multi-trillion edge graphs on millions of cores in seconds. In *Proceedings of 2018 ACM Intern. Conf. on Supercomputing*.
2. Morris, G., Goodsell, D., Halliday, R.S., Huey, R., Hart, W., Belew, R., and Olson, A. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 19 (1998), 1639–1662.
3. Zhang, Y., Wang, H., Jia, M., Wang, J., Li, D.-S., Xue, G. and Tan, K. TopoX: Topology refactorization for minimizing network communication in graph computations. *IEEE/ACM Trans. Networking* 28 (2020), 2768–2782.

**Yiming Zhang** is a professor at National University of Defense Technology in Changsha, China.

**Kai Lu** is a professor at National University of Defense Technology in Changsha, China.

**Wenguang Chen** is a professor at Tsinghua University in Beijing, China.

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

Aggregated GPEPS

	#Vtx	#Edge	Size	#Node	IO cost(s)	PR	BFS	WCC	BC	KCore	Net. Usage
twitter	41.6m	1.4b	16GB	16	37.2	0.75	0.36	0.54	0.30	0.01	28.1%
uk-2007	105.8m	3.7b	41GB	64	28.9	3.99	3.36	3.97	0.87	0.13	28.1%
clueweb	978.4m	42.5b	476GB	100	119.4	5.04	2.47	4.79	1.23	0.41	47.4%
uk-2014	747.8m	47.6b	532GB	100	132.7	4.67	2.06	3.81	0.35	0.32	47.1%
weibo	349.9m	42.4b	474GB	100	121.3	8.22	2.56	7.15	4.77	0.72	40.8%
hyberlink	3.5b	128.7b	1.5TB	256	243.0	17.20	3.68	14.49	0.32	0.26	60.1%
sogou	271.8b	12253.9b	136.9TB	38656	2130.6	1443.4	30.8	224.4			18.8%
34.rmat	17.1b	274.8b	3.3TB	1024	-	83.0	34.5	77.7	44.8	0.25	29.2%
34.erdos	17.1b	274.8b	3.3TB	1024	-	52.5	48.8	51.3	40.9	18.60	63.7%
34.kron	17.1b	274.8b	3.3TB	1024	-	72.8	18.2	62.3	43.1	0.82	40.2%
42.kron	4398.0b	70368.7b	844.4TB	38656	-	1984.8	773.9	1956.0			40.0%

Dataset and performance of Shentu.

BY GUOFENG ZHANG, XIAOWEI ZHOU, FENG TIAN,  
HONGBIN ZHA, YONGTIAN WANG, AND HUJUN BAO

# The Present and Future of Mixed Reality in China

VIRTUAL REALITY (VR) technology can be used to generate a lifelike virtual world, allowing users to roam freely in this virtual world and interact with it. Augmented reality (AR) is a technology that superimposes and fuses virtual scenery or information with the real physical environment, and interactively presents it in front of users, so that the virtuality and reality share the same space. Mixed reality (MR) is a further development of VR and AR. Its concept was proposed by Paul Milgram and Fumio

Kishino in 1994, who described the entire continuum from reality to virtuality.<sup>14</sup>

VR/AR research in China began in the late 1990s. In 2002, the 973 project (National Basic Research Program) “Fundamental Theories, Algorithms of Virtual Reality and Its Implementation” was established, with Hujun Bao from Zhejiang University as the chief scientist. In





2009, the project received continuous support, and the focus was shifted to AR and MR. The research of these two projects lasted for 10 years. Zhejiang University, the Institute of Software and the Institute of Automation of the Chinese Academy of Sciences, Beijing Institute of Technology (BIT), Beihang University, Tsinghua University, and other institutions participated in these

two projects. The projects made many world-class research advances in VR/AR related areas such as automatic camera tracking, efficient modeling, real-time rendering, multichannel human-computer interaction, VR/AR headset display, 3D registration, as well as VR/AR engines and software platforms. The projects also trained a group of outstanding VR/AR

researchers, which significantly boosted VR/AR research in China.

With the continuous investment of national funding agencies and research institutions, the quality of China's MR research has developed rapidly. The number of publications from Chinese researchers in top conferences such as IEEE VR and the International Symposium on Mixed and Augmented Reality

(ISMAR) surged. For example, the first-author affiliations of 18.5% of accepted papers for ISMAR 2020, a top conference of AR and MR, are from China, rising from only 4% in 2010. Moreover, researchers from SenseTime and Zhejiang University received the best paper award of ISMAR 2020 for their work Mobile-3DRecon,<sup>25</sup> which was the first time authors from Chinese institutions won this award. With the increasing research impact, more and more Chinese researchers have been invited to join the organization committees of top conferences. In 2019, ISMAR was held in China for the first time. Qingping Zhao from Beihang University and Yongtian Wang from BIT served as general chairs, and Shi-Min Hu from Tsinghua University served as the science and technology chair.

Meanwhile, the nation has made further plans for the MR industry. In December 2018, the Ministry of Industry and Information Technology of China issued “Guiding Opinions on Accelerating the Development of the Virtual Reality Industry.” It proposed establishing “a relatively complete virtual reality industry chain” in China by 2020, and achieving the goal of “the overall strength of China’s virtual reality industry will be in the forefront of the world by 2025.” In March 2021, China’s 14<sup>th</sup> Five-Year Plan was officially announced, listing VR and AR as key industries in the digital economy over the next five years.

### Research Contributions from Chinese Scholars in MR

To realize the immersive visual fusion and presence of virtual and real environments, key technologies such as tracking and registration, 3D modeling, realistic rendering, human-computer interaction, and natural display need to be developed. Next, we will introduce the current state of research in these key technologies in China.

**3D registration and reconstruction.** For a mixed reality system, 3D registration technology mainly aims to reconstruct the 3D information of the real scene and the real-time pose information of the user or the

camera. Simultaneous localization and mapping (SLAM) is a most important 3D registration technique, which realizes the inside-out tracking and localization. Although the research on SLAM in China started late, it has also made remarkable achievements. For example, the RDSLAM<sup>16</sup> and RKSLAM<sup>13</sup> proposed by Bao’s team at Zhejiang University are both well-known monocular visual SLAM systems. The former can handle dynamic environments, and the latter can run on a smartphone in real-time. The VINS-Mono<sup>15</sup> proposed by Shaojiao Shen’s team from the Hong Kong University of Science and Technology (HKUST) has become one of the most popular open-source visual-inertial SLAM systems due to its remarkable robustness and versatility. ICE-BA,<sup>12</sup> which was jointly developed by Baidu and Zhejiang University, increases the speed of bundle adjustment (BA) by an order of magnitude using the incremental computation. Some Chinese companies have also successively launched SLAM-based AR/MR platforms, for example, SenseTime’s SenseAR<sup>a</sup> and Huawei’s AR Engine.<sup>b</sup>

With the progress of deep learning in recent years, some deep learning-based SLAM methods have also emerged. However, these methods typically require collection of large-scale training data, and are generally difficult to generalize to environments that have changed or have never been seen before. In order to solve this issue, the team of Hongbin Zha from Peking University proposed a SLAM framework based on the online learning paradigm, which enables the SLAM system to infer uncertainty and quickly adapt itself in a rapidly changing environment.<sup>11,22</sup>

In the area of 3D reconstruction, both academia and industry in China made great progress. The ENFT-SfM<sup>28</sup> proposed by Zhejiang University achieves fast and robust large-scale scene reconstruction through an efficient non-consecu-

tive feature tracking method and segment-based bundle adjustment algorithm. The HSfM<sup>2</sup> proposed by the Institute of Automation of the Chinese Academy of Sciences achieves similar accuracy to the incremental reconstruction through a hybrid structure-from-motion method, and at the same time reaches an efficiency close to or even better than that of the global reconstruction. The ultra-large-scale global 3D reconstruction system proposed by HKUST realizes a distributed global motion averaging through a divide-and-conquer approach and achieves efficient 3D reconstruction from tens of thousands of images.<sup>30</sup> In industry, a series of 3D reconstruction products have emerged, such as Altizure, DJI Terra, AirlookMap, and others, which can automatically reconstruct high-precision 3D models from image data taken by drones.

### Simulation and rendering.

Computer simulation has been developed for two decades in China. In the 1990s, Jiaoying Shi, Qunsheng Peng, and Enhua Wu started their pioneering research, covering considerable subjects in AR/VR. Since then, Chinese scholars have been devoting greater effort and making significant contributions. For complex elasticity, Bao’s team has proposed a series of technologies (for example, Huang et al.<sup>6</sup>) to greatly improve the simulation efficiency. For fluids, Enhua Wu, Guoping Wang, Shi-Min Hu, and Xiaopei Liu and their teams proposed novel methods (for example, Yan et al.<sup>23</sup>) to efficiently simulate complex fluid phenomena including multi-phase fluids. Kun Zhou and his team focused on human bodies and solved fidelity problems in hair and face simulation.<sup>1</sup> The teams of Qingping Zhao, Xiaogang Jin, Min Tang, and Mingliang Xu proposed to leverage the ideas of space-time continuity to address problems in medical simulation, group animation, and cloth simulation. Wang et al.<sup>17</sup> developed haptic devices and algorithms in VR, overcoming the problems caused by high frame rates.

In generalized real-time rendering, significant achievements have

a <http://openar.sensetime.com/>

b <https://developer.huawei.com/consumer/en/hms/huawei-arengine/>

been seen in stereo rendering, automatic shader optimization, and global illumination. Stereo and power-aware rendering is an important research direction to provide realistic content for head-mounted displays, which have limited computational resources. Bao's team developed a cutting-edge stereo shading and shader optimization architecture,<sup>26</sup> which is promising to meet the ever-increasing requirements of high-framerate and high-resolution for immersive VR. Neural rendering is a prominent direction for providing premium effects in a uniform framework. Zhejiang University, Tsinghua University, and Kujiale contributed leading solutions in this domain, ranging from real-time single-bounce indirect illumination, denoising, to path guiding.<sup>20</sup> For the rendering of specific material and effects, the teams from Zhejiang University and Nanjing University introduced state-of-the-art real-time techniques<sup>21,4</sup> for rendering cloth and participating media, respectively.

#### Human-computer interaction.

Human-computer interaction is a key component in MR research, which involves various aspects including theories, devices, and techniques. In the area of theory, the team of Feng Tian from the Institute of Software, Chinese Academy of Sciences (ISCAS) proposed an uncertainty model based on the ternary Gaussian probability distribution.<sup>7,8</sup> It can accurately predict the target acquisition error rate and has an important guiding role in the interaction designs in MR. By combining touch-screen gestures and a variety of tactile feedback mechanisms based on electrostatic force, Zhao et al.<sup>29</sup> designed a multi-channel visual- and touch-based virtual reality interactive system for 3D object interaction tasks, which improves the accuracy of 3D object operations.

In the research of interactive devices, the team of Yingqing Xu from Tsinghua University made a breakthrough in the 1:N scanning drive and push-push contact latch technology, which effectively solves the problem of large-format render-

ing with tactile dot-matrix feedback devices, and effectively improves the resolution and stability. Wang et al.<sup>19</sup> developed wearable tactile feedback gloves that can achieve 0-4N continuous and stable force feedback.

In the research of interactive techniques, the team of Yongtian Wang from BIT proposed a calibration method based on a dynamic pinhole camera model to solve the problem of precise virtual-real fusion in close-range high-precision hand-eye collaborative interaction. ISCAS designed an interactive component (vMirror) for interactions in VR, which uses the reflection of the mirror to observe and select long-distance occluded objects, improve the selection efficiency of occluded targets, and reduce user dizziness.<sup>10</sup>


#### Optical display of VR/AR devices.

Head-mounted display (HMD) is an important carrier device of VR and AR. HMD for VR (VR-HMD) has been deployed in commercial applications earlier than AR-HMD, and the technical solution has become increasingly mature.


Due to disadvantages like small field of view (FOV), inferior image quality, and heavy design, the applications of early AR-HMD with off-axis mirror and relay lens are limited. The introduction of freeform surfaces not only greatly increases the design freedom, but also significantly reduces the volume and weight of AR-HMDs. Zhejiang University, Nankai University, and BIT conducted in-depth research on freeform optics. The team of Yongtian Wang from BIT proposed a closed-loop optimization design method that integrates full-FOV image quality balance and injection error pre-compensation. The freeform element they developed weighs only 8g, which greatly reduces the weight of an AR-HMD.<sup>3</sup> In 2018, Ned+ announced the boundless AR optics module with freeform optics,<sup>c</sup> which has a diagonal FOV of 120°.


Chinese researchers have also done excellent work to further reduce the volume and weight of AR-HMD.

c <http://www.nedplusar.com/en/index>




**Terminal-cloud collaboration through 5G networks will make it possible to achieve the high-resolution, frame rate, and fidelity rendering of large-scale virtual scenes on mobile or head-mounted MR devices.**





**Stereo and power-aware rendering is an important research direction to provide realistic content for head-mounted displays, which have limited computational resources.**



In 2015, the team of Qiang Sun from Changchun Institute of Optics and Mechanics proposed a waveguide structure with two vertically arranged half-reflective films and achieved a FOV of  $20^{\circ} \times 15^{\circ}$  and two-dimensional expansion of the exit pupil.<sup>9</sup> In 2020, based on a dual-layer geometrical waveguide, Wang et al.<sup>18</sup> from BIT proposed an ultra-thin, large-FOV AR-HMD which achieved a total thickness of 3.0 mm,  $62^{\circ}$  FOV, and 10-mm exit pupil at an eye relief of 18 mm.

An AR-HMD based on diffractive optical elements has also been developed. In 2011, Yan et al.<sup>24</sup> proposed a method of dispersion-free diffraction using four holographic gratings optimization method for holographic waveguide, which achieved a circular FOV of  $25^{\circ}$  and a large pupil of about 43 mm. In 2015, Han et al.<sup>5</sup> from BIT designed a waveguide display system composed of freeform elements and volume holographic gratings with a diffraction efficiency of 87.57%, which achieves a diagonal FOV of  $45^{\circ}$ . In 2017, based on space-variant volume holographic gratings, Chao et al.<sup>27</sup> proposed a quite efficient waveguide display which achieved 31.9% system efficiency,  $20^{\circ}$  FOV, and high brightness uniformity simultaneously.

### **The Mixed Reality Industry in China**

In recent years, many VR/AR startup companies have emerged in China. Quite a few of them are focused on VR/AR all-in-one systems or core component devices. For example, Pico, QiYu VR, NOLO, and others have launched VR all-in-one systems. Shadow Creator, Nreal, and more have launched birdbath-based AR glasses. Collaborating with Ned+, BIT launched a variety of AR-HMD products based on freeform surface optics. Ned+, Lochn Optics, Lingxi, Rokid, Greater, North Ocean Photonics and others have launched waveguide AR-HMD. In recent years, Chinese start-up companies appeared at the Consumer Electronics Show and the Augmented Reality World Expo and won a series of awards.

In addition to hardware, a few tech giants in China have also launched AR/MR technology platforms. Baidu released its open platform DuMix AR<sup>d</sup> in 2017, empowering developers with AR technology. Huawei released the AR Engine for Huawei mobile phones in 2018 and its spatial computing platform Cyberspace in 2019. SenseTime released the SenseAR developer platform in 2018 and upgraded it in 2020 to SenseMARS, a cross-hardware and cross-system MR platform providing high-precision 3D mapping and spatial computing capabilities. Compared with the international leading products, some of China's MR products (such as SenseMARS and Cyberspace) already can achieve almost comparable MR effects in large-scale scenes, and even have some differentiated advantages in some respects. For example, to the best of our knowledge, the MR platform SenseMARS jointly developed by SenseTime and Zhejiang University is the first product in the industry that is able to achieve high-precision 6DoF visual-inertial localization and AR navigation in large-scale indoor scenes on the Web and mini programs.

The development of MR in China is at a relatively early stage and mainly based on scenarios involving camera apps, short video, and live broadcasting for applications such as house viewing and navigation guides. In 2015, AR effects selfie camera app "FaceU" was released and quickly became popular and was acquired in 2018 by ByteDance for about US\$300 million. Douyin and Kuaishou have added AR special effects to their short video applications. In the real estate sector, Beike took the lead in launching the VR house viewing function in 2018. In e-commerce, Alibaba and JD.com launched the buy+ and Tiangong virtual shopping plans in 2016 and 2017 respectively to provide e-commerce with digital AR content, though they are not yet open to large-scale third-party users. In the past two years, more and more companies, including Didi,

<sup>d</sup> <https://dumix.baidu.com/>


Huawei, SenseTime, and Baidu, have launched visual localization with AR navigation solutions or applications.

### Future Perspectives

Although there is still a certain gap between China's MR products and international leaders, the gap is narrowing rapidly. In particular, with the popularization of 5G and the rapid development of cloud computing in China, high-bandwidth and low-latency networks will greatly push MR technologies towards the combination of terminal and cloud computing, and city-level MR will be realized in the near future. For example, SLAM technology can be combined with high-precision 3D maps and cloud computing to break through the bottleneck of robustness and efficiency in large-scale scenes. Also, the terminal-cloud collaboration through 5G networks will make it possible to achieve the high-resolution, frame rate, and fidelity rendering of large-scale virtual scenes on mobile or head-mounted MR devices. Zhejiang University has made outstanding contributions in the related research.

In addition, key technologies in MR, such as rendering, simulation, 3D modeling, and interaction are also being deeply integrated with AI in China. It is expected that in the next few years, China's MR products will not only work in a much larger environment but will also become more high-fidelity and in effect smarter. For instance, neural scene representation and rendering show promise to break through the limitations of traditional graphics pipelines. Tsinghua University, Microsoft Research Lab Asia, and Zhejiang University have made remarkable contributions. Moreover, current MR systems still lack sufficient intelligence to allow group participation and collaboration. Deep integration with AI is required. Some well-known companies in China—for example, Huawei and SenseTime—have devoted huge effort to MR+AI research and products.

Last but not least, the sense of

immersion, ease of use, and wearing comfort of MR equipment are essential for commercialization of MR. Optical display technology is the key. The future optical solutions of VR-HMD will mainly focus on aspheric lens VR-HMD, Fresnel lens VR-HMD, and pancake VR-HMD. AR-HMD is more challenging. Free-form AR-HMD can expand the field of view with high image quality, and commercial mass production is possible in China. Geometric waveguides and diffractive waveguides can achieve a thinner system structure. However, the former is difficult to mass produce. With China's progress in high-precision optical component processing and design, diffractive optical waveguides, in terms of optical effects, appearance, and mass production prospects, will become the mainstream of AR-HMD in China. 

### References

- Cao, C., Weng, Y., Lin, S., Zhou, K. 3D shape regression for real-time facial animation. *ACM Trans. Graphics* 32, 4 (2013), 1–10.
- Cui, H., Gao, X., Shen, S., Hu, Z. HSfM: Hybrid structure-from-motion. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 2017, 1212–1221.
- Cheng, D., Wang, Y., Hua, H., Talha, M. Design of an optical see-through head-mounted display with a low f-number and large field of view using a freeform prism. *Applied Optics* 48, 14 (2009), 2655–2668.
- Guo, J., Guo, Y., Pan, J., Lu, W. BRDF analysis with directional statistics and its applications. *IEEE Trans. Visualization and Computer Graphics* 26, 3 (2020), 1476–1489.
- Han, J., Liu, J., Yao X., Wang, Y. Portable waveguide display system with a large field of view by integrating freeform elements and volume holograms. *Optics Express* 23, 3 (2015), 3534–3549.
- Huang, J., Tong, Y., Zhou, K., Bao, H., Desbrun, M. Interactive Shape Interpolation through Controllable Dynamic Deformation. *IEEE Trans. Visualization and Computer Graphics* 17, 7 (2011), 983–992.
- Huang, J., Tian, F., Fan, X., Zhang, X., Zhai, S. Understanding the uncertainty in 1D unidirectional moving target selection. In *Proceedings of the ACM Conf. Human Factors in Computing Systems*, 2018, 1–12.
- Huang, J., Tian, F., Fan, X., Tu, H., Zhang, H., Peng, X., Wang, H. Modeling the endpoint uncertainty in crossing-based moving target selection. In *Proceedings of the 2020 CHI Conf. Human Factors in Computing Systems*, 2020, 1–12.
- Hu, X., Sun, Q., Li, J., Li, C., Liu, L., Zhang, J. Spectral dispersion modeling of virtually imaged phased array by using angular spectrum of plane waves. *Optics Express* 23, 1 (2015), 1–12.
- Li, N., Zhang, Z., Liu, C., Yang, Z., Fu, Y., Tian, F., Han, T., Fan, M. vMirror: enhancing the interaction with occluded or distant objects in VR with virtual mirrors. In *Proceedings of 2021 CHI Conf. Human Factors in Computing Systems*, 1–11.
- Li, S., Wang, X., Cao, Y., Xue, F., Yan, Z., Zha, H. Self-supervised deep visual odometry with online adaptation. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 2020, 6339–6348.
- Liu, H., Chen, M., Zhang, G., Bao, H., Bao, Y. ICE-BA: incremental, consistent and efficient bundle adjustment for visual-inertial SLAM. In *Proceedings of the IEEE Conf. Computer Vision and Pattern Recognition*, 2018, 1974–1982.
- Liu, H., Zhang, G., Bao, H. Robust keyframe-based monocular SLAM for augmented reality. In *Proceedings of the IEEE Intern. Symp. Mixed and Augmented Reality*, 2016, 1–10.
- Milgram, P., Kishino, F. A taxonomy of mixed reality visual displays. *IEICE Trans. Information Systems* E77-D 12, (1994), 1321–1329.
- Qin, T., Li, P., Shen, S. VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Trans. Robotics* 34, 4 (2018), 1004–1020.
- Tan, W., Liu, H., Dong, Z., Zhang, G., Bao, H. Robust monocular SLAM in dynamic environments. In *Proceedings of the IEEE Intern. Symp. on Mixed and Augmented Reality*, 2013, 209–218.
- Wang, D., Zhang, Y., Hou, J., Wang, Y., Lv, P., Chen, Y., Zhao, H. iDental: A haptic-based dental simulator and its preliminary user evaluation. *IEEE Trans. Haptics* 5, 4 (2011), 332–343.
- Wang, Q., Cheng, D., Hou, Q., Gu, L., and Wang, Y. Design of an ultra-thin, wide-angle, stray-light-free near-eye display with a dual-layer geometrical waveguide. *Optics Express* 28, 23 (2020), 35376–35394.
- Wang, Z., Wang, D., Zhang, Y., Liu, J., Wen, L., Xu, W., Zhang, Y. A three-fingered force feedback glove using fiber reinforced soft bending actuators. *IEEE Trans. Industrial Electronics* 67, 9 (2019), 7681–7690.
- Xu, B., Zhang, J., Wang, R., Xu, K., Yang, Y., Li, C., Tang, R. Adversarial Monte Carlo denoising with conditioned auxiliary feature modulation. *ACM Trans. Graphics* 38, 6 (2019), 224:1–224:12.
- Xu, C., Wang, R., Zhao, S., Bao, H. Multi-scale hybrid micro-appearance modeling and realtime rendering of thin fabrics. *IEEE Trans. Visualization and Computer Graphics* 27, 4 (2019), 2409–2420.
- Xue, F., Wang, X., Wang, J., Zha, H. Deep visual odometry with adaptive memory. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2020.
- Yan, X., Jiang, Y., Li, C., Martin, R., Hu, S. Multiphase SPH simulation for interactive fluids and solids. *ACM Trans. Graphics* 35, 4 (2016), 1–11.
- Yan, Z., Li, W., Zhou, Y., Kang, M., Zheng, Z. Virtual display design using waveguide hologram in conical mounting configuration. *Optical Engineering* 50, 9 (2011) 94001.
- Yang, X., Zhou, L., Jiang, H., Tang, Z., Wang, Y., Bao, H., Zhang, G. Mobile3DRecon: real-time monocular 3D reconstruction on a mobile phone. *IEEE Trans. Visualization and Computer Graphics* 26, 12 (2020), 3446–3456.
- Yuan, Y., Wang, R., Bao, H. Tile pair-based adaptive multi-rate stereo shading. *IEEE Trans. Visualization and Computer Graphics* 26, 6 (2020), 2303–2314.
- Yu, C., Peng, Y., Zhao, Q., Li, H., and Liu, X. Highly efficient waveguide display with space-variant volume holographic gratings. *Applied Optics* 56, 34 (2017), 9390–9397.
- Zhang, G., Liu, H., Dong, Z., Jia, J., Wong, T., Bao, H. Efficient non-consecutive feature tracking for robust structure-from-motion. *IEEE Trans. Image Processing* 25, 12 (2016), 5957–5970.
- Zhao, L., Liu, Y., Ye, D., Ma, Z., Song, W. Implementation and evaluation of touch-based interaction using electrovibration haptic feedback in virtual environments. In *Proceedings of the 2020 IEEE Conf. Virtual Reality and 3D User Interfaces*, 239–247.
- Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L. Very large-scale global SfM by distributed motion averaging. In *Proceedings of the 2018 IEEE Conf. Computer Vision and Pattern Recognition*, 4568–4577.

**Guofeng Zhang** is a professor at Zhejiang University, Hangzhou, China.

**Xiaowei Zhou** is a research professor at Zhejiang University, Hangzhou, China.

**Feng Tian** is a professor at the Institute of Software, Chinese Academy of Sciences, Beijing, China.

**Hongbin Zha** is a professor at Peking University, Beijing, China.

**Yongtian Wang** is a professor at Beijing Institute of Technology, Beijing, China.

**Hujun Bao** is a professor at Zhejiang University, Hangzhou, China.

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

BY CHUN YU AND JIAJUN BU

# The Practice of Applying AI to Benefit Visually Impaired People in China

ACCORDING TO THE China Disabled Persons' Federation (CDPF), there are now 17 million visually impaired people in China, among which three million are totally blind, while the others are low-visioned. In the past two decades, China has experienced tremendous development of information technology. Traditional industries are incorporating information technology, with services delivered to users through websites and mobile applications. It is positive technical progress that visually impaired people can access various services without leaving home; for

example, they can order food delivery online or schedule a taxi from an app-based transportation service.

However, the development of technology has also brought challenges to the visually impaired in China. First, the cost to make massive information services barrier-free is huge. Information accessibility per se is challenging due to visual impairment coupled with IT developers' poor awareness of information accessibility. These factors result in a large portion of applications that do not meet accessibility standards. Second, the development of technology has led to urbanization and a fast pace of life, and the outdoor environment is not suitable for the visually impaired to walk alone. It is also challenging to develop technology that enables visually impaired people to walk in complex outdoor environments. The development of artificial intelligence creates the opportunity to address these challenges.

CDPF works to establish and promote China's own standard system of information accessibility. The joint force combines the power of the government, universities, and enterprises like Baidu, Alibaba, and Tencent, among others. Zhejiang University, as a member of the Federation, has taken the lead in formulating China's first national Internet information accessibility standard. There are four main principles that provide the foundation for this standard: Perceivability, Operability, Understandability, and Robustness. The standard incorporates 58 standard terms for website and mobile application accessibility, which are divided into three levels based on their influence on barrier-free use, universality and scalability, and the difficulty of technical implementation. This standard can guide Internet content providers to gradually improve their accessible service capabilities. This national standard is being promoted in coordination with the World Wide Web Consortium's Web Content Accessibility Guidelines (WCAG) 2.1, and China has advertised the standard as





**The Taobao app is making shopping more accessible to visually impaired users.**

“tactile paving on the Internet.”

In spite of those efforts, it is still challenging to meet the accessibility standard when developing Internet products, due to the lack of accessibility awareness of developers, inadequate understanding of the real needs of users, and the inability to simulate real user behaviors. Taking advantage of artificial intelligence, media computing, and crowdsourcing technologies, Zhejiang University in Hangzhou, China, has assembled a substantial body of research on URL-clustering-based Web page sampling algorithms and active learning-based sampling algorithms,<sup>11,12</sup> the barrier point detec-

tion method,<sup>7</sup> an automatic evaluation system based on Web Accessibility Evaluation Metric (WAEM) barrier weights, evaluation task classification and scheduling algorithms<sup>1</sup> (shown in Figure 1), user experience prediction algorithms,<sup>6</sup> barrier weight optimization algorithms based on user feedback,<sup>5</sup> large-scale data analysis, and so forth. Figure 1 provides an overview of the task classification and scheduling algorithms, which utilize historical user data to train a model and make a correlation analysis after clustering. The result is an assignment map based on which tasks can be assigned to evaluators and how the evaluation

results can be analyzed.

Figure 2 shows the overall process of this crowdsourcing-based Internet information accessibility evaluation system. The system achieves a higher accessibility evaluation accuracy with a lower labor cost and is more in line with the real user feelings of visually impaired users. Involving users' real feedback in the process can help analyze the impact of different detection items on users' intuitive experience and help the evaluation result match users' real experience as much as possible. Since 2012, more than 2,000 Chinese government websites have been evaluated annually, includ-

Figure 1. Task classification and scheduling algorithms overview.

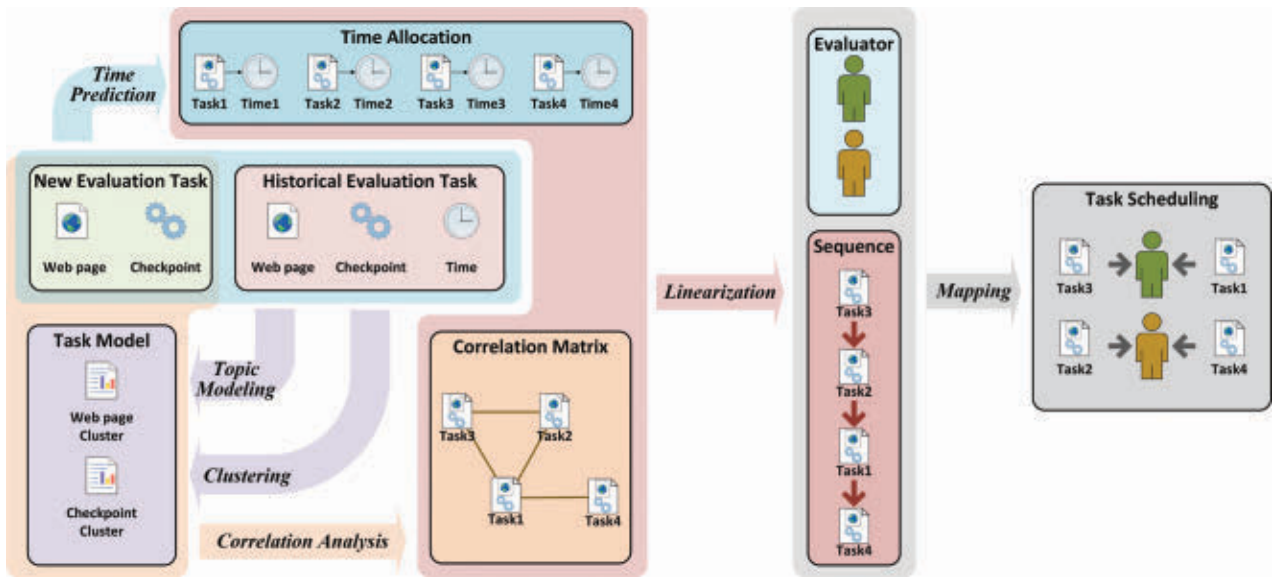


Figure 2. Zhejiang University's crowdsourcing-based Internet information accessibility evaluation system.

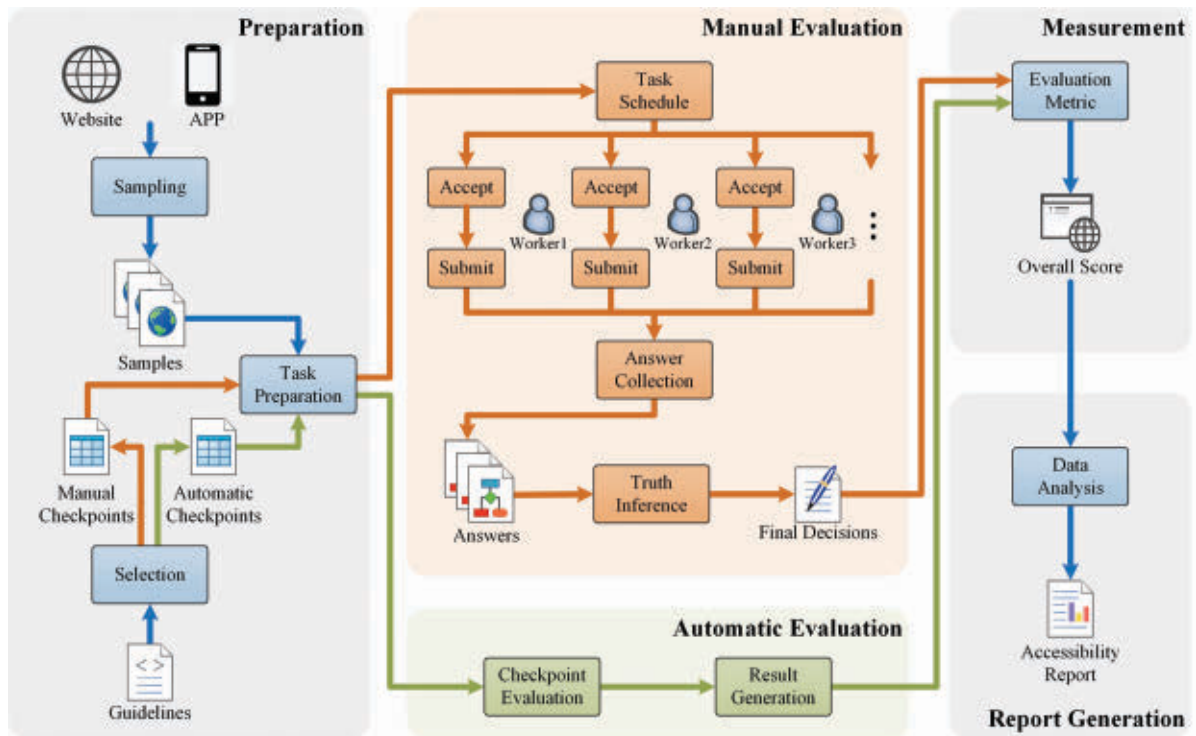


Figure 3. (Left) VIPBoard; (Right) Eartouch.

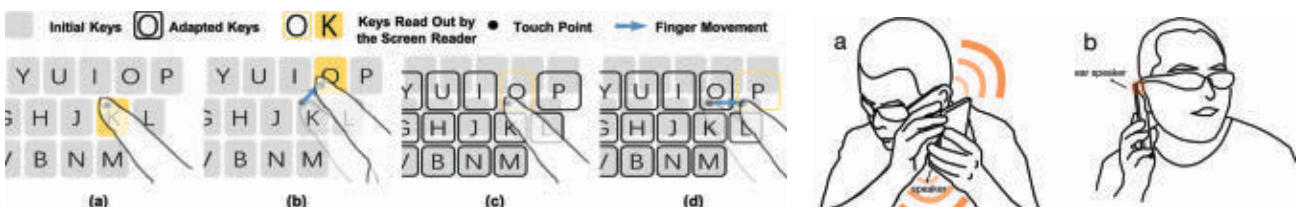
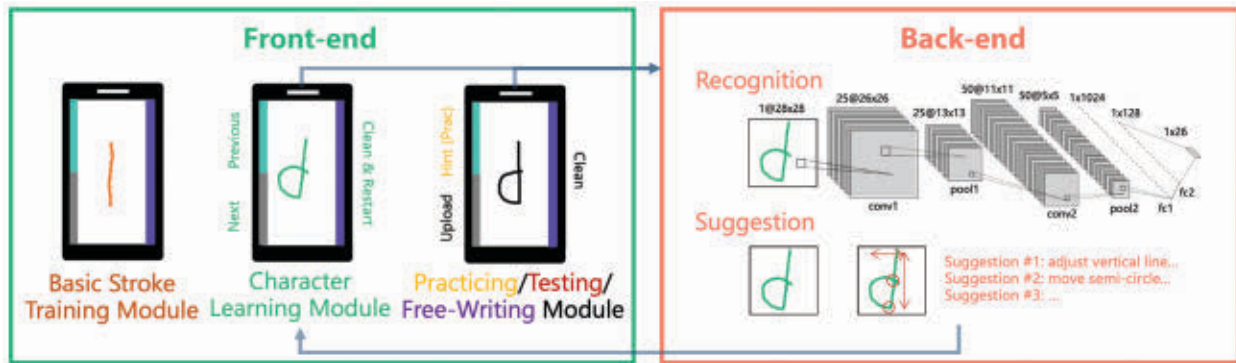


Figure 4. The architecture of LightWrite system.



ing the ministries, commissions, offices, and agencies directly under the state council, state bureaus administered by ministries or commissions, and provincial government websites in China, all promoting the Chinese government's e-service quality.

### Innovating Interaction Techniques for Information Accessibility

Information accessibility is a popular research area in human-computer interaction. In recent years, with the rapid development of sensing and computing technology, researchers have explored ways to break through the GUI paradigm for accessible use, innovating intelligent and higher levels of barrier-free experience. Tsinghua University represents an activist toward this direction in China; in particular, its research efforts highlight a systematic intersection of identifying user need by consulting schools for the blind, innovating interaction techniques, and deploying them into practice by collaborating with IT companies.

A good example is the VIPBoard,<sup>3</sup> an intelligent keyboard technique designed for visually impaired users. Visually impaired users rely on audio feedback to interact with a smartphone. This makes the word-level autocorrection algorithm of modern software keyboards unusable, because a user cannot proceed to type until hearing the wanted letter. The researchers applied their experience on intelligent input to solve this problem. They iterated a series of solutions, and finally came up with a character-level error-correction mechanism, which eliminates up to 65% of corrections and improved text entry by 11%. Then,

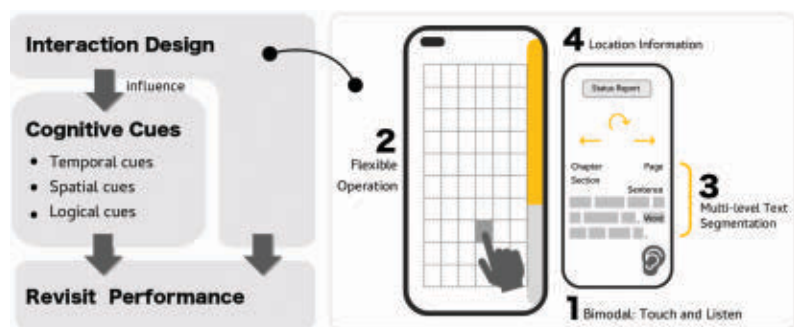
via collaboration with Sogou Inc., the VIPBoard technique was integrated into the largest input method software in China and is now serving thousands of users daily. Similarly, to improve the usability of smartphones, the researchers propose EarTouch,<sup>8</sup> which leverages a capacitive screen to recognize and locate a user's ear in contact with the screen. EarTouch enables users to input content with one hand and can protect a user's privacy in public environments. EarTouch has been integrated into Smart Screen Reader, which not only benefits thousands of visually impaired users, but also serves as a platform to experiment with new ideas and innovations regarding information accessibility. VIPBoard and EarTouch (as illustrated in Figure 3) both won Honorable Mention Awards at ACM CHI 2019, the leading conference in the field of human-computer interaction.


Beyond addressing the basic input requirement of smartphones, elevating their cultural level is crucial for visually impaired people to have better job opportunities and improve their quality of life. Researchers in Tsinghua University

recognize the breakthrough point to be innovating low-cost and easy-to-access techniques, so each visually impaired user can reap the benefit. LightWrite<sup>9</sup> is an AI teacher on smartphones that uses voice-based descriptive instruction and feedback to teach visually impaired users to write English letters and Arabian digits in a specifically designed stroke. It can teach users handwritten characters, with an average of 1.09 minutes required for each letter. LightWrite serves as a practical solution for teaching writing (see Figure 4).


To enable extensive reading, researchers focused on providing support for revisitation, which is the essential skill of comparing concepts and improving understanding. A variety of navigation gestures and multimodal feedbacks were designed and tested. The final reading interface provides multiple spatial and temporal cues so users can locate the content they have read quickly. Lab experiments showed that an app-based reader with multiple feedbacks could achieve a high level of accuracy and efficiency for revisitation in reading and outperformed the hard-

Figure 5. Revisitation model and design of the reader.





**It is still challenging to meet the accessibility standard when developing Internet products, due to the lack of accessibility awareness of developers, inadequate understanding of the real needs of users, and the inability to simulate real user behaviors.**



ware point display reader that costs thousands of dollars. Both techniques significantly reduce the cost for visually impaired users to improve their level of culture (see Figure 5).

### **Applications of Accessibility Technologies**

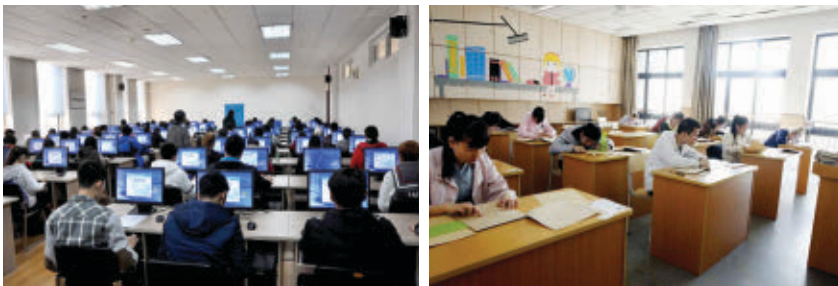
In cooperation with Alibaba, Zhejiang University researched related technologies, such as reading order optimization and image structure understanding, to help visually impaired users obtain image information. The graph-to-sequence-based end-to-end reading order technology incorporated with the OCR-based image structure learning algorithm has been applied in the screen reader developed by CDPF. Zhejiang University also participated in the development of a detection platform for Alibaba to explore rapid solutions for Internet content, such as computer and mobile terminals. A number of Internet commercial services, such as the Taobao online shopping platform and the Alipay online payment platform, have undergone barrier-free transformation in accordance with the national standards. They have optimized 37 functions of the Taobao App, covering basic services such as login and registration, product search, product purchase, receipt confirmation, and rights protection. According to incomplete statistics, there were more than 160,000 online shops on Taobao capable of use by people with disabilities, and 2.46 million people with disabilities used them to make purchases (see Figure 6).


The China Braille Library, Alibaba groups, and Zhejiang University have set up an example of a room in an accessible smart home. The control center mainly consists of a smart speaker designed by Tmall (formerly TaoBao Mall) connecting with more than 30 smart home hardware products such as sensors, robot vacuums, and smart TVs. It improves accessibility in the security, cleaning, illumination, amusement, circular control, and kitchen areas; thus, the visually impaired can control household electrical devices via voice and realize accessible living (as shown in Figure 7). The smart office hardware enterprise represented by Alibaba groups has established an accessibility alliance on smart office hardware. The schools for

the blind in 31 provinces of China have specified and deployed a batch of smart office hardware to introduce facial recognition and collaborative working technologies to the special education field. During the COVID-19 pandemic, it guaranteed teaching activities would continue normally and made communication with blind students and their parents accessible, realizing smart accessible school management.

China has increasingly developed special education in the 21<sup>st</sup> century, growing from two universities established before 2000 to 18 universities today, with many still making preparations. For example, Changchun University first proposed integrating special education into normal higher education, so that other than teaching on professional courses, students in a special educational college could obtain the same cultural quality education, public elective courses, recreational and sports activities and matches, as fully sighted students. Inclusive education is not only conducive to eliminating the unhealthy mentality of disabled students, such as fear of intimacy, and feelings of inferiority and paranoia, but also helps disabled students come in contact with cutting-edge information technology and enter first-class Internet enterprises (see Figure 8).

The China Braille Press and the Institute of Computing Technology of the Chinese Academy of Sciences have designed automated technology for two-way translation between written Chinese and braille. Traditional translation methods require a large amount of manual checking and amending, while the new translation technology combines the N-Gram language model with the rule of phrase translation and creates an improved language model, which can not only get rid of invalid homonym word strings according to braille word segmentation, but also allows full phrases in context to be translated into braille. During the translation, the new technology makes use of the tones of Chinese braille to reduce some mismatched candidates among Chinese characters. The new technology, which can attain 91.43% accuracy when translating Chinese to braille, and 90.32% accuracy when translating braille to Chinese, can be applied in real-world applications such

**Figure 6. Visually impaired users shopping online.****Figure 7. An example room of an accessible smart home.****Figure 8. The classroom for special education.**


technologies. Companies in industry understand how the Internet and markets work, in the context of the development of a new assistive product and a complete system for bringing it to market, while the government can make use of and transfer the technology into products to bring them to market using mature market operation mechanisms to push products and services to end users that need them. In addition, China has many special user groups such as the Disabled Persons' Federation and the Blind Persons' Federation, which can help to assure the products and services can satisfy user requirements. 

#### References

- Li, L., Wang, C., Song, S., Yu, Z., Zhou, F., and Bu, J. A task assignment strategy for crowdsourcing-based Web accessibility evaluation system. In *Proceedings of the 14th Web for All Conf. on The Future of Accessible Work*. (2017).
- Liu, G., Xu, H., Yu, C., Xu, H., Xu, S., Yang, C., Wang, F., Mi, H., and Shi, Y. Tactile compass: Enabling visually impaired people to follow a path with continuous directional feedback. In *Proceedings from CHI 2021*, 1–13.
- Shi, W., Yu, C., Fan, S., Wang, F., Wang, T., Yi, X., Bi, X., and Shi, Y. VIPBoard: Improving screen-reader keyboard for visually impaired people with character-level auto correction. In *Proceedings from CHI 2019*, 517.
- Song, S., Bu, J., Artmeier, A., Shi, K., Wang, Y., Yu, Z., and Wang, C. Crowdsourcing-based Web accessibility evaluation with golden maximum likelihood inference. In *Proceedings of the 2018 ACM on Human-Computer Interaction, CSCW*, 1–21.
- Song, S., Bu, J., Wang, Y., Yu, Z., Artmeier, A., Dai, L., and Wang, C. Web accessibility evaluation in a crowdsourcing-based system with expertise-based decision strategy. In *Proceedings of the 2018 Internet of Accessible Things*.
- Song, S., Bu, J., Shen, C., Artmeier, A., Yu, Z., and Zhou, Q. Reliability aware web accessibility experience metric. In *Proceedings of the 2018 Internet of Accessible Things*.
- Song, S., Wang, C., Li, L., Yu, Z., Lin, X., and Bu, J. WAEM: a web accessibility evaluation metric based on partial user experience order. In *Proceedings of the 14th Web for All Conf. The Future of Accessible Work*. (2017).
- Wang, R., Yu, C., Yang, X., He, W., and Shi, Y. EarTouch: Facilitating Smartphone use for visually impaired people in mobile and public scenarios. In *Proceedings from CHI 2019*, 24.
- Wu, Z., Yu, C., Xu, X., Wei, T., Zou, T., Wang, R., and Shi, Y. LightWrite: Teach handwriting to the visually impaired with only a smartphone. *Proceedings from CHI 2021*, 1–15.
- Xu, S., Yang, C., Ge, W., Yu, C., and Shi, Y. Virtual Paving: Rendering a smooth path for people with visual impairment through vibrotactile and audio feedback. In *Proceedings of ACM Interact. Mob. Wearable Ubiquitous Technology* 4, 4 (2020), 99:1–99:25.
- Yu, Z., Bu, J., Shen, C., Wang, W., Dai, L., Zhou, Q., and Zhao, C. A multi-site collaborative sampling for Web accessibility evaluation. In *Proceedings of the Intern. Conf. Computers Helping People with Special Needs*. Springer, Cham, 2020, 329–335.
- Zhang, M., Wang, C., Bu, J., Yu, Z., Lu, Y., Zhang, R., and Chen, C. An optimal sampling method for Web accessibility quantitative metric. In *Proceedings of the 12th Web for All Conf.* (2015).

**Chun Yu** is an associate professor at Tsinghua University in Beijing, P.R. China.

**Jiajun Bu** is a professor at Zhejiang University in Hangzhou, P.R. China.

 This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License. <https://creativecommons.org/licenses/by-nd/4.0/>

as editing and publishing braille books and establishing braille instructional materials.

#### Prospect

China's unrestricted technology is in the stage of rapid development, characterized by the combination of innovation and practice, and supported by the China Disabled Persons' Federation and technology giants. Through the innovation and traction of universities, such technology can be put on the ground as soon as possible. In the future, more efforts will be put

into making technology barrier-free, not only to support the blind, but also to better support the elderly and other groups with special needs.

The AI can help visually impaired people integrate into society and obtain information on an equal basis. In this article, we have combined the power of the government, industry, and academia. The government, in charge of establishing policy, rules, and management systems, plays a leading role, taking advantage of experts and technology from universities and research institutes to improve the key

BY ZHIWEN YU, HUADONG MA,  
BIN GUO, AND ZHENG YANG

## Crowdsensing 2.0

MOBILE CROWDSENSING (MCS) presents a new sensing paradigm based on the power of user-companioned devices.<sup>11,12</sup> It allows “the increasing number of smartphone users to share local knowledge acquired by their sensor-enhanced devices, and the information can be further aggregated in the cloud for large-scale sensing.”<sup>4</sup> The mobility of large-scale mobile users makes MCS a versatile platform that can often replace static sensing infrastructures. A broad range of applications are thus enabled, including traffic planning, environment monitoring, urban management, and so on.

During the past decade, MCS has become a surging research topic in China. There are several reasons that precede this new sensing paradigm:

- ▶ *Widespread new techniques.* With the rapid advancement of mobile communication (4G/5G) and pervasive sensing techniques, it has been reported that sensor-rich smartphone users reached around 800 million in China by 2020, forming the largest ‘mobile’ population around the world. The prevalence of mobile devices in China builds a solid physical foundation for crowdsensing.

- ▶ *Promotion by national research and development plans.* China has put significant efforts into improving



MCS through a series of key projects under its national R&D plans. These plans cover a series of major techniques and application areas, such as the Internet of Things (IoT), smart cities, as well as the next generation of artificial intelligence (AI 2.0<sup>13</sup>).

- ▶ *Particular developmental challenges and opportunities.* As a developing country, China is undergoing a large-scale urbanization process. Numerous complex challenges have been raised, such as environment protection, transportation optimization, and urban management. The emergence of MCS opens up new opportunities to address these challenges.

The term “mobile crowdsensing” was coined in Ganti et al.’s work<sup>14</sup> in 2011. In 2012, Yunhao Liu from Tsinghua University gave a thorough characterization of the definition for the first time, noting research



challenges as well as MCS opportunities.<sup>11</sup> Huadong Ma et al. from Beijing University of Posts and Telecommunications (BUPT) proposed several new metrics (for example, opportunistic coverage) to characterize the sensing opportunity and quality.<sup>12</sup> The team led by Zhiwen Yu at Northwestern Polytechnical University (NPU) gave a thorough review of the challenges, the architecture, novel applications of MCS, and proposed the Visual Crowdsensing (VCS) concept,<sup>3,4</sup> which leverages built-in cameras of smartphones to attain informative/visual sensing of interesting targets. Tsinghua University and Shanghai Jiaotong University (SJTU) systematically study the general incentive mechanisms in MCS,<sup>6,20</sup> as it is important to have the active participation of citizens. Zhang et al. from Peking University investigated Sparse Mobile Crowdsensing,<sup>16</sup> which lever-

ages the spatial and temporal correlation among the data sensed to reduce the required number of sensing tasks allocated.

#### Key applications and beyond.

Besides scientific investigation, many novel MCS applications (as shown in Figure 1) have also been developed in China to address the societal and developmental issues in different domains.

► *Environment protection.* Rapid urbanization usually results in severe air pollution problems (for example, PM2.5), especially for cities in developing countries like China. However, until now it has been difficult to obtain fine-grained citywide PM2.5 status due to insufficient monitoring sites. Third-Eye,<sup>9</sup> a crowdsensing application developed for fine-grained PM2.5 monitoring, is rooted in a joint research by BUPT and Microsoft Research Asia. It uses

advanced deep learning algorithms to infer PM2.5 levels simply using the outdoor images taken by citizens' smartphones. This helps government and researchers collect fine-grained data to better understand the causes and the propagation mechanism of PM2.5 pollution in modern cities. It has been integrated in Microsoft's Urban Air<sup>a</sup> product.

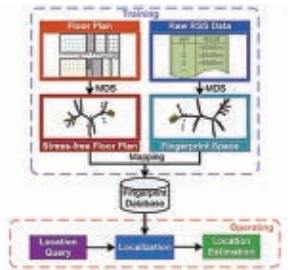
► *Indoor localization.* Existing indoor localization techniques are mostly based on radio-based solutions, which usually require a site survey process to obtain detailed radio signals of interested areas. However, site survey proves costly in time and manpower, limiting its usage in practice. Researchers from Tsinghua University designed LiFS,<sup>17</sup> which is a crowdsensing application

a <https://www.microsoft.com/en-us/research/project/urban-air/>

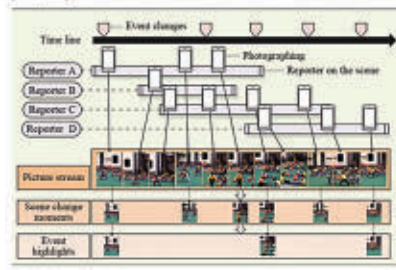
Figure 1. Representative MCS apps.



(a) The ThirdEye and Urban Air project, a joint research from BUPT and MSRA



(b) The LiFS project from Tsinghua



(c) The InstantSense project from NPU

that leverages smartphones to construct the radio map of floor plans.

► *Refined urban management.* Smart city development relies on urban and community dynamics monitoring to provide essential information. Google launched the Waze service,<sup>b</sup> a successful MCS application in the traffic system domain. It allows drivers to share roadside events (for example, road work, traffic jams, accidents) and aggregate them on the digital map for navigation. In 2021, Didi Chuxing,<sup>c</sup> the leading mobile transportation company in China, launched a new application called “Long-distance Eye,” which allows citizens to share visual information of roadside events (for example, photos or short videos about traffic accidents) through their smartphones or dashcams. This allows for better driving plans and reduces traffic congestion. Digital China<sup>d</sup> is a famous smart city service provider. In 2013, they released the “Integrated Citizen Service Platform,” which allows citizens to instantly report various urban management issues they encounter (for example, pavement/sidewalk damage). A similar application is the SeeClickFix<sup>e</sup> in the U.S., which allows people to report non-emergency neighborhood issues to local governments.

► *Event sensing* in cities is crucial for identifying emergency/unusual events and maintaining public safety. Researchers from NPU in China have developed InstantSense,<sup>1</sup> which leverages peoples’ photographs to obtain detailed event reports in real time.

**Toward Heterogeneous Crowdsensing**

MCS is a human-centered sensing paradigm, but some places are not accessible for human beings. In recent years, with the development of IoT and edge computing techniques, there are more pervasive devices equipped with sensing and computing capabilities. With the integration of these capabilities, the next-generation MCS (MCS 2.0)—Heterogeneous Crowdsensing (HCS)—is proposed. There are several distinctive features that characterize

b <https://www.waze.com/>  
 c <https://www.didiglobal.com/>  
 d <http://www.digitalchina.com/en/>  
 e <https://seeclickfix.com/>

Figure 2. Heterogeneous crowdsensing (MCS 2.0).

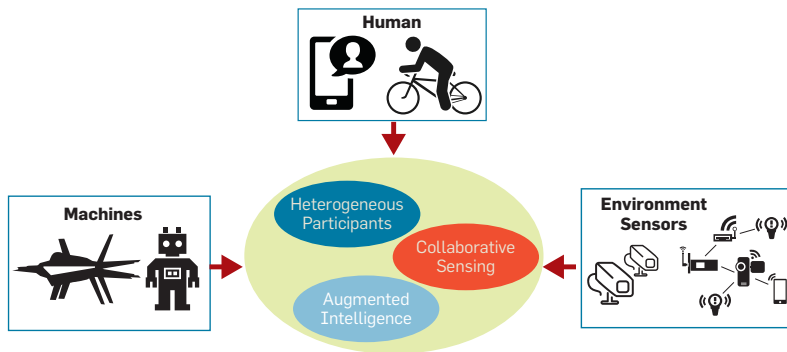
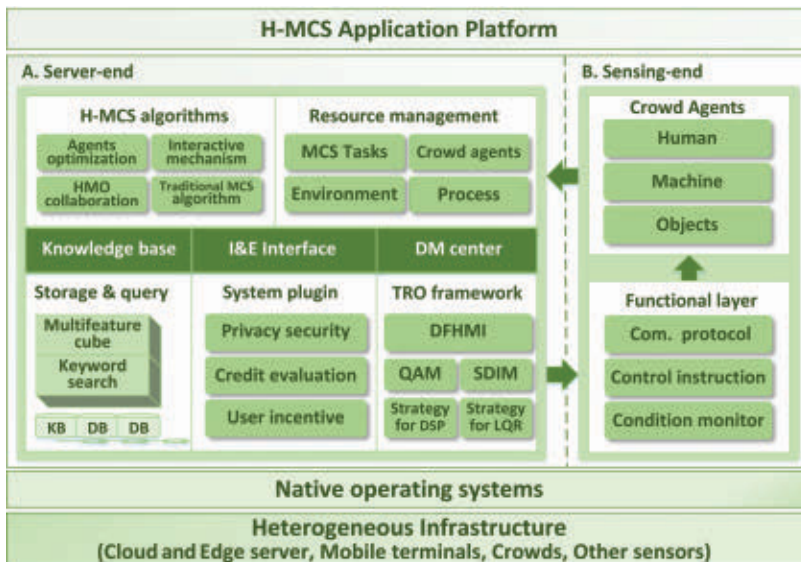


Figure 3. CrowdOS 2.0 kernel.





HCS (as shown in Figure 2).

**Heterogeneous participants.** HCS aims to build an extended sensing and computing space with heterogeneous participants, including humans, machines, and environmental sensors. The *pervasive machines* involve various robots, smart vehicles, and so on, while environmental sensors consists of sensor networks.

**Collaborative sensing.** The sensing capabilities of different participants are distinct. Though humans are skilled in mobile sensing, environmental sensors have advantages for constant sensing in sparsely populated regions. HCS studies collaborative sensing techniques to adaptively select and aggregate the complementary sensing power from diverse participants to complete complex sensing tasks.

**Augmented intelligence.** A distinctive feature of HCS is that both human and machine participation is involved in the large-scale sensing process. The coexistence of human intelligence and machine intelligence, however, must be orchestrated in an appropriate and optimal manner to enhance both.

### Enabling Techniques

MCS is heading toward the HCS generation, and there are several enabling techniques that help achieve this goal.

**Energy-efficient collaborative sensing.** MCS 2.0 leverages pervasive devices as additional participants for crowdsensing. However, energy concerns for such resource-constrained devices are emerging regarding their limited power supply and high computational needs.<sup>15</sup> To address this issue, collaborative sensing becomes an important technique. There are currently two major methods under investigation. The first method eliminates the dependency of on-device batteries by mingling heterogeneous energy-harvesting mechanisms to harvest energy from surrounding objects.<sup>2</sup> The second way is to replace traditional energy-consuming sensors with energy-efficient sensors regarding the spatio-temporal dependencies.<sup>8</sup>

**Hybrid human-machine intelligence (HHMI).** In 2017, the Chinese government issued the AI 2.0 plan.<sup>13</sup> According to the plan, breakthroughs should be made in several


basic theories of AI, including the hybrid human-machine intelligence (HHMI). HHMI aims to integrate human intelligence into an AI system to complement machine capabilities throughout its life cycle. However, the implementation of HHMI is still being explored. First, a set of theory is required for defining human and machine functions and characteristics. Second, criteria should be set to estimate the opportunity of human-machine cooperation during the continuous learning cycle. Third, the evaluation methods of the HHMI system from the perspective of both humans and machines is also an important but difficult issue. As a representative case of HHMI, a human-feedback identification model is proposed,<sup>19</sup> which can continuously update the tree-based incremental learning model with human guidance.

**Operating system for crowdsensing.** Regarding the vast crowdsourcing field, Amazon Mechanical Turk<sup>f</sup> has become the most successful online crowdsourcing company in the world. The core of it is a generic framework that supports various online crowdsourcing tasks (for example, translation and image labeling). Different from online crowdsourcing, MCS represent unique features such as spatio-temporality and pervasive sensing. To support rapid design and development of MCS applications in different domains, numerous platforms and frameworks are emerging. Existing MCS platforms usually address specific issues from a certain perspective, and different platforms are not compatible with each other. In 2019, researchers from NPU in China addressed this issue by drawing on the idea of ubiquitous operating systems (OS) and proposed a novel OS (CrowdOS 1.0),<sup>10</sup> which is an abstract software layer running between the native OS and the application layer. Based on an in-depth analysis of the complex relationship among crowdsensing tasks, participants, and data, they built the OS kernel with three core modules: the Task Resolution and Assignment Framework (TRAF), the Integrated Resource

<sup>f</sup> <https://www.mturk.com/>



**HCS aims to build an extended sensing and computing space with heterogeneous participants, including humans, machines, and environmental sensors.**



Management module (IRM), and the Task Result Quality Optimization (TRO) module.

To meet the new characteristics of MCS 2.0, a new version of CrowdOS is being developed, to deal with the new challenges such as heterogeneous resource management, scheduling, and collaboration, as shown in Figure 3. The 2.0 version further enhances the platform's compatibility and extensibility for heterogeneous MCS. As an open source MCS platform, CrowdOS shares the application installation packages, related algorithm modules, source codes, and various project documents from its website ([www.crowdos.cn](http://www.crowdos.cn)). To date, there have been more than 20,000 visits from 20+ countries. Based on CrowdOS, we are now collaborating with researchers from around the world (for example, Rutgers University in the U.S., Ulster University in the U.K., Waseda University in Japan) and Trustie<sup>g</sup> (the biggest open source software community in China) to launch an open source competition on innovative MCS applications. This is believed to be the first competition on open source MCS software development over the world.

### Future Trends

Though we have witnessed significant development of MCS in China in recent years, there are still several issues to be tackled to further promote its growth in the MCS 2.0 era.

**Strengthen the theoretical foundation.** Though MCS has received much achievements at the technological level, the theoretical foundation of it is largely lagging behind. The basic scientific issues, such as the emergence of crowd intelligence (that is, the wisdom of crowds), crowd cognition mechanisms, and hybrid human-machine intelligence, are still not sufficiently explored. Therefore, one fundamental task is to study these theoretical problems and using the mechanisms/principles to improve MCS technique development.<sup>7</sup>

**Embracing new technologies.** The future of MCS must be aware of new technologies and applications. The emergence of AI, AIoT (AI in IoT), and

Blockchain brings new opportunities and benefits to MCS. For instance, AIoT leverages compression and edge-intelligence methods to enable the usage of deep learning algorithms in resource-constrained devices.<sup>21</sup> Blockchain, however, paves the way for trust and secure interaction among independent agents.<sup>5,18</sup>


### Deployment of killer apps of MCS.

MCS is still limited to a supporting role in China and not placed in a dominate position to deal with major developmental issues. By aggregating the benefits of human, machine, and IoT devices, the next generation of MCS will present grand opportunities in key national economical areas such as smart industry and social governance. Therefore, the Chinese MCS community should maintain close collaboration with industry researchers and government managers to investigate killer MCS apps to be deployed.

### Conclusion

Mobile crowdsensing has rapidly developed in China during the last decade. The pervasive use of smart/wearable devices as well as the large-scale sensing requirements from different fields drives its development. Numerous enabling techniques are explored or being developed, including task allocation and worker selection, incentive mechanisms, crowdsourced data management, privacy/security protection, as well as collaborative sensing and human-machine intelligence. The availability of a generic crowdsensing platform such as CrowdOS also facilitates the research and development of MCS applications. Recently, the Chinese government released the "Fourteenth Five Year Plan and the 2035 Vision of China," where the development of "Digital Economics" and "Digital China" has become a core mission. MCS is expected to play an important role in the coming digital economics era.

### Acknowledgment

This work was partially supported by the National Key R&D Program of China (2019YFB2102200), and the National Science Fund for Distinguished Young Scholars (61725205, 62025205). 

### References

- Chen, H., Guo, B., Yu, Z. and Han, Q. Toward real-time and cooperative mobile visual sensing and sharing. *IEEE INFOCOM 2016*, 1–9.
- Cui, X., Zhang, J., Zhou, H. and Deng, C. PowerPool: Multi-source ambient energy harvesting. In *Proceedings of the 8th Intern. Conf. Big Data Computing and Communications*, 2020, 86–90.
- Guo, B., Han, Q., Chen, H., Shangquan, L., Zhou, Z. and Yu, Z. The emergence of visual crowdsensing: Challenges and opportunities. *IEEE Commun. Surveys & Tutorials* 19, 4 (2017), 2526–2543.
- Guo B, Wang Z, Yu Z, Wang Y, Yen N, Huang R, Zhou X. Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. *ACM Computing Surveys* 48, 1 (2015), 1–31.
- Huang, J. et al. Blockchain-based mobile crowd sensing in industrial systems. *IEEE Trans. Industrial Informatics* 16, 10 (2020), 6553–6563.
- Jin, H., Guo, H., Su, L., Nahrstedt, K., Wang, X. Dynamic task pricing in multi-requester mobile crowd sensing with Markov correlated equilibrium. *IEEE INFOCOM 2019*, 1063–1071.
- Li, W., Wu, W., Wang, H., Cheng, X., Chen, H., Zhou, Z. and Ding, R. Crowd intelligence in AI 2.0 era. *Frontiers of Information Technology & Electronic Engineering* 18, 1 (2017), 15–43.
- Liang, Y., Wang, X., Yu, Z., Guo, B., Zheng, X. and Samtani, S. Energy-efficient collaborative sensing: Learning the latent correlations of heterogeneous sensors. *ACM Trans. Sensor Networks*, 2021.
- Liu L, Liu W, Zheng Y, Ma H, and Zhang C. Third-Eye: A mobilephone-enabled crowdsensing system for air quality monitoring. In *Proceeding of the ACM Interact. Mob. Wearable Ubiquitous Technol.* vol. 2, no. 1, 2018, 26 pages.
- Liu, Y., Yu, Z., Guo, B., Han, Q., Su, J., and Liao, J. CrowdOS: A ubiquitous operating system for crowdsourcing and mobile crowd sensing. *IEEE Trans. Mobile Computing*, 2021.
- Liu, Y. Crowd sensing and computing. *Commun. CCF* 8, 10 (2012), 38–41.
- Ma, H., Zhao, D., Yuan, P. Opportunities in mobile crowd sensing. *IEEE Commun. Mag.* 52, 8 (2014), 29–35.
- Pan, Y. Heading toward artificial intelligence 2.0. *Engineering* 2, 4 (2016), 409–413.
- Raghu, K., Ganti, F., Ye, H., Lei, L. Mobile crowdsensing: current state and future challenges. *IEEE Commun. Mag.* 49, 11 (2011), 32–39.
- Wang, J., Wang, Y., Zhang, D. and Helal, S. Energy saving techniques in mobile crowd sensing: Current state and future opportunities. *IEEE Commun. Mag.* 56, 5 (2018), 164–169.
- Wang, L., Zhang, D., Wang, Y., Chen, C., Han, X. and M'hamed, A. Sparse mobile crowdsensing: Challenges and opportunities. *IEEE Commun. Mag.* 54, 7 (2016), 161–167.
- Wu, C., Yang, Z., Liu, Y. Smartphones based crowdsourcing for indoor localization. *IEEE Trans. Mobile Computing* 14, 2 (2015), 444–457.
- Xu, J., Wang, S., Bhargava, B. and Yang, F. A blockchain-enabled trustless crowd-intelligence ecosystem on mobile edge computing. *IEEE Trans. Industrial Informatics* 15, 6 (2019), 3538–3547.
- Yang, F., Yu, Z., Chen, L., Gu, J., Li, Q., and Guo, B. Human-machine cooperative video anomaly detection. In *Proceedings of the 23rd ACM Conf. Computer-Supported Cooperative Work and Social Computing*, 2020, 1–18.
- Zhang, X., Yang, Z., Sun, W., Liu, Y., Tang, S., Xing, K., Mao, X. Incentives for mobile crowd sensing: A survey. *IEEE Commun. Surveys & Tutorials* 18, 1 (2016), 54–67.
- Zhou, Z., Liao, H., Gu, B., Huq, K., Mumtaz Sand Rodriguez, J. Robust mobile crowd sensing: When deep learning meets edge computing. *IEEE Network* 32, 4 (2018), 54–60.

Zhiwen Yu is a professor at Northwestern Polytechnical University, Xi'an, China.

Huadong Ma is a professor at Beijing University of Posts and Telecommunications, Beijing, China.

Bin Guo is a professor at Northwestern Polytechnical University, Xi'an, China.

Zheng Yang is a professor at Tsinghua University, Beijing, China.

<sup>g</sup> <https://www.trustie.net/>

**Several companies are trying push automatic speech recognition and other technologies past their current limitations.**

**BY JIA JIA, WEI CHEN, KAI YU, XIAODONG HE, JUN DU, AND HEUNG-YEUNG SHUM**

# The Practice of Speech and Language Processing in China

ALTHOUGH GREAT PROGRESS has been made in automatic speech recognition (ASR), significant performance degradation still exists in very noisy environments. Over the past few years, Chinese startup AISpeech has been developing very deep convolutional neural networks (VDCNN),<sup>21</sup> a new architecture the company recently

began applying to ASR use cases.

Different than traditional deep CNN models for computer vision, VDCNN features novel filter designs, pooling operations, input feature map selection, and padding strategies, all of which lead to more accurate and robust ASR performance. Moreover, VDCNN is further extended with adaptation, which can significantly alleviate the mismatch between training and testing.

Factor-aware training and cluster-adaptive training are explored to fully utilize the environmental variety and quickly adapt model parameters. With this newly proposed approach, ASR systems can improve the system robustness and accuracy, even in under very noisy and complex conditions.<sup>1</sup>

JD AI Research (JD), based in Beijing, China, has also made progress in auditory perception,

Figure 1. Models for sound event detection and localization.

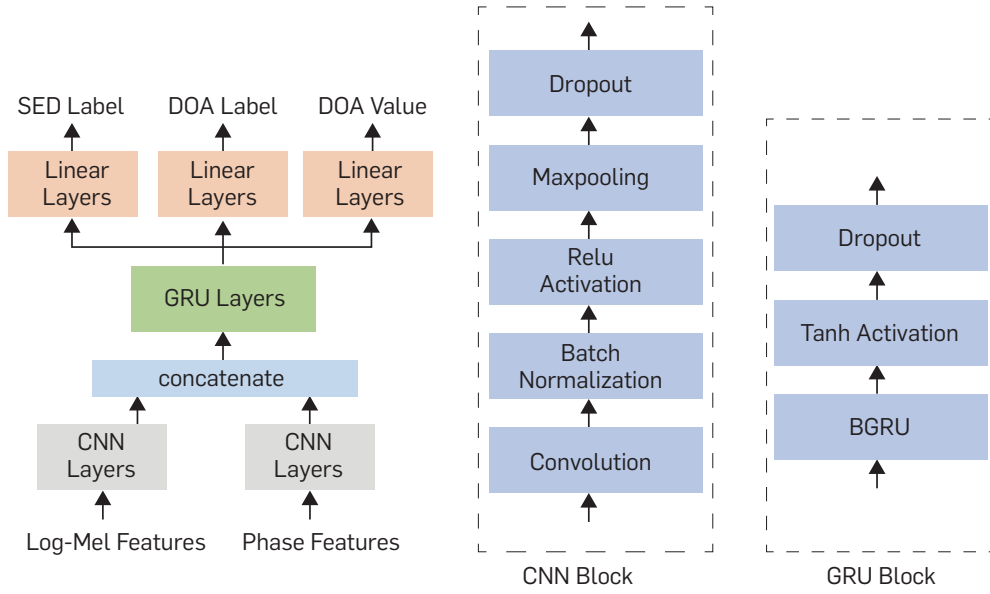


Figure 2. System diagram of densely connected multi-stage model for real-time speech enhancement.

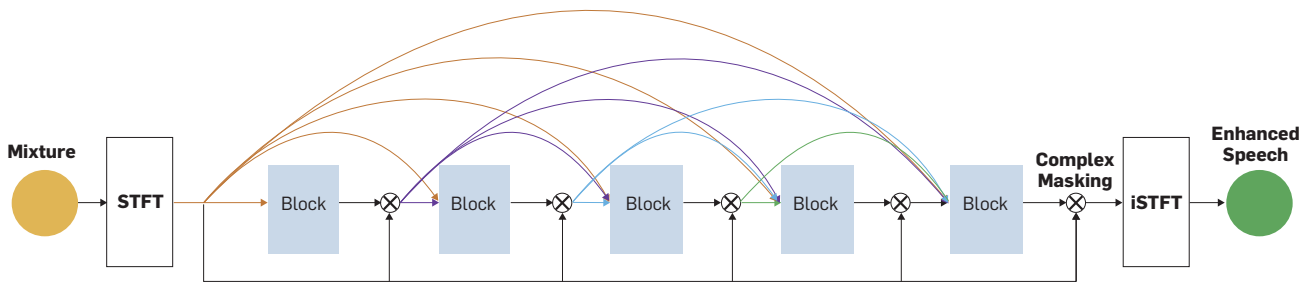
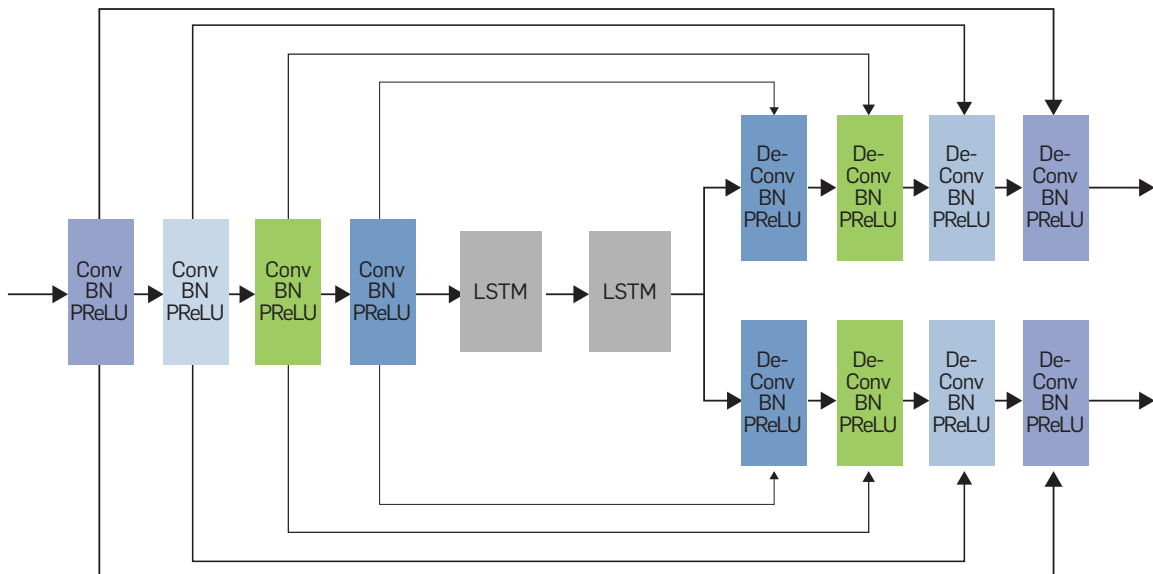


Figure 3. Block of the system.



aiming to detect and localize sound events, enhance target signals, and suppress reverberation. This is important not only because it enhances signals for speech recognition, but also because such information can be used for better decision-making in subsequent dialog systems.

For sound-event detection, as shown in Figure 1, a multi-beamforming-based approach is proposed: the diversified spatial information for the neural network is extracted using beamforming towards different directions.<sup>32</sup> For speech dereverberation, optimal smoothing-factor-based preprocessing is used to obtain a better presentation for the dereverberation network.<sup>10</sup> Beamforming and speech dereverberation are also used to generate augmented data for multichannel far-field speaker verification.<sup>22</sup> In terms of speech enhancement, neural Kalman filtering (KF) is proposed to combine conventional KF and speech evolution in an end-to-end framework.<sup>31</sup>

JD also ranked third in both the sound event localization and detection task of DCASE 2019 Challenge, and the FFSVC 2020 Challenge for far-field speaker verification.

For real-time speech enhancement, China-based Internet company Sogou proposes a deep complex convolution recurrent network (DC-CRN) with restricted parameters and latency.<sup>9</sup> Different from real-valued

networks, DCCRN adopts the complex CNN, complex long short-term memory (LSTM), and complex batch normalization layers, which are better suited for processing complex-valued spectrograms. Moreover, as shown in Figure 2 and Figure 3, a computational, efficient, real-time speech-enhancement network is proposed with densely connected, multistage structures.<sup>11</sup> The model applies sub-band decomposition and progressive strategy to achieve superior denoising performance with lower latency.

For end-to-end ASR, self-attention networks (SAN) in transformer-based architectures<sup>23</sup> show promising performance, so a transformer-based, attention-based encoder/decoder (AED) is selected as the base architecture.

One approach is to improve AED performance for non-real-time speech transcription. Transformer-based architectures can easily achieve slightly better results than traditional hybrid systems in ordinary scenarios. However, transformer-based models collapse under some conditions, such as conversational speech and recognition of proper nouns. Relative positional embedding (RPE) and parallel scheduled sampling (PSS)<sup>39</sup> are adopted to improve generalization and stability. As transformer architecture is good at global modeling, and speech recognition relies more on local information, local

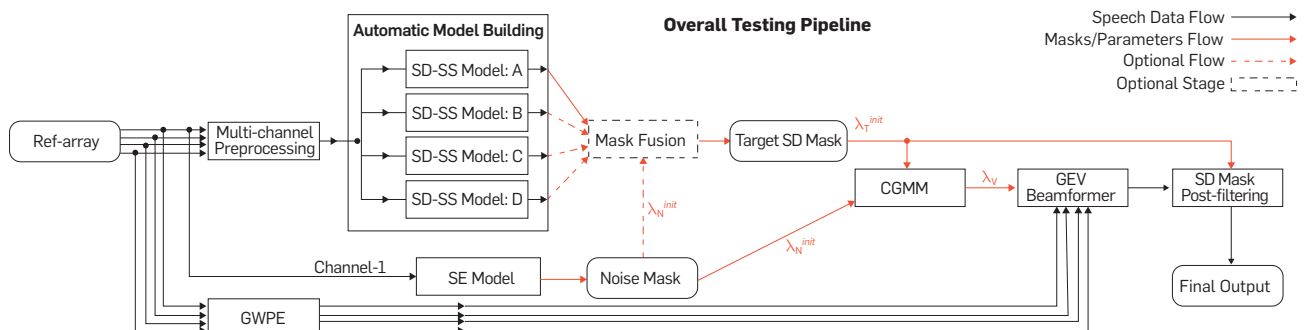
modeling is further combined with CCNs and feedforward sequential memory networks (FSMN)<sup>7</sup> to the transformer to improve the modeling of local speech variance. To improve acoustic feature extraction of encoders, Sogou uses connectionist temporal classification (CTC) and cross entropy (CE), multitask joint training of the transformer. With this strategy, a 100,000-hour transformer achieves a 25% improvement compared to Kaldi-based hybrid systems.

A second research strategy is streaming AED. To that end, Sogou proposed an adaptive monotonic chunk-wise attention (AMoChA) mechanism,<sup>6</sup> which can adaptively learn chunk-length at each step to calculate context vectors for streaming attention. Transformer acoustic range is adaptively computed for each token in a streaming decoding fashion. For the CTC and CE joint-trained transformer, CTC output is viewed as first-pass decoding while the attention-based decoder is seen as second-pass decoding. Thus, the encoder is trained in a chunk-wise manner for streaming AED. This method is similar to non-auto-regressive decoding.<sup>8</sup>

The 100,000-hour streaming AED achieved a 15%–20% relative improvement compared to Kaldi-based hybrid streaming systems. Generally, ASR systems and speech enhancement (SE) systems are trained and deployed separately, because

**Figure 4. The overall diagram of USTC-iFLYTEK front-end processing system for the CHIME-5 challenge.<sup>20</sup>**

Here, SD-SS means speaker-dependent speech separation, SE model is a deep learning-based speech enhancement model, GWPE denotes the generalized weighted prediction error algorithm for dereverberation, CGMM means complex Gaussian mixture model, and GEV means generalized eigenvalue.



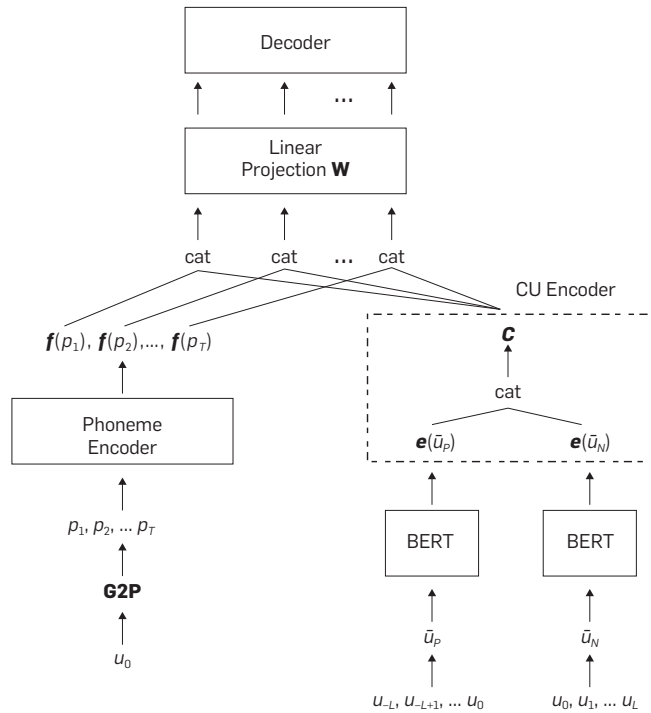
**DCCRN adopts the complex CNN, complex long short-term memory (LSTM), and complex batch normalization layers, which are better suited for processing complex-valued spectrograms.**

they typically have different purposes. Moreover, enhanced speech is detrimental to ASR performance. However, joint training of SE and ASR can significantly improve the performance of speech in high-noise environments while maintaining the performance of clean speech. For Sogou, the joint training system of the CRN-based SE model and the transformer-based ASR model results in an average relative improve-

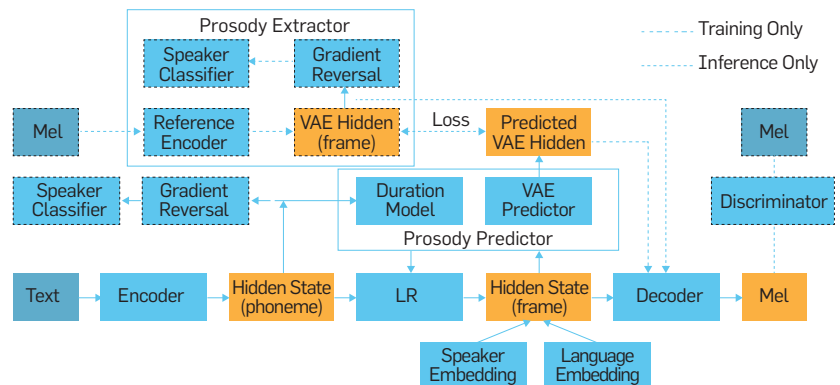
ment of 23% in noisy conditions and 5% in clean conditions.

Visual information is another way to boost speech recognition performance in noisy conditions. Google first proposed the Watch, Listen, Attend and Spell (WLAS) network, which jointly learns audio and visual information in the recognition task.<sup>4</sup> Sogou adopted a modality attention network based on WLAS<sup>40</sup> for adaptively integrating audio and visual

**Figure 5. The embeddings for the future and past chunked sentences are concatenated to form the Cross Utterance (CU) context vector, which is concatenated with the phoneme encoder output vectors to form the input of the decoder.**



**Figure 6. StyleTTS architecture.**



information, which achieved a 35% performance improvement in 0-dB noisy conditions.

iFLYTEK, together with the National Engineering Laboratory for Speech and Language Information Processing at the University of Science and Technology of China (USTC), proposed novel, high-dimensional regression approaches to solve classical speech-signal preprocessing problems and is outperforming traditional methods by relaxing the constraints of many mathematical model assumptions.<sup>5,20,29</sup> The organization has finished in first place in several prestigious challenges, including all four tasks of the CHiME-5 speech recognition challenge,<sup>20</sup> two tasks of the CHiME-6 speech recognition challenge,<sup>27</sup> all tasks of the DIHARD-III Speech Diarization Challenge,<sup>15</sup> and the Sound Event Localization and Detection (SELD) task of the DCASE2020 Challenge.<sup>13</sup> These challenges, especially CHiME-5/6 and DIHARD-III, are quite relevant to common “cocktail party problems” found in real multi-speaker scenarios. Figure 4 shows an overview of the USTC-iFLYTEK front-end processing system for the CHiME-5 challenge.

### Robust Speaker Identification

Deep learning-based methods have been widely applied in this research area, achieving a new milestone for speaker identification and anti-

spoofing. However, it is still difficult to develop a robust speaker identification system under complex, real-world scenarios such as short utterance, noise corruption, and channel mismatch. To boost speaker verification performance, AISpeech proposes new approaches to achieve more discriminant speaker embeddings within two frameworks.

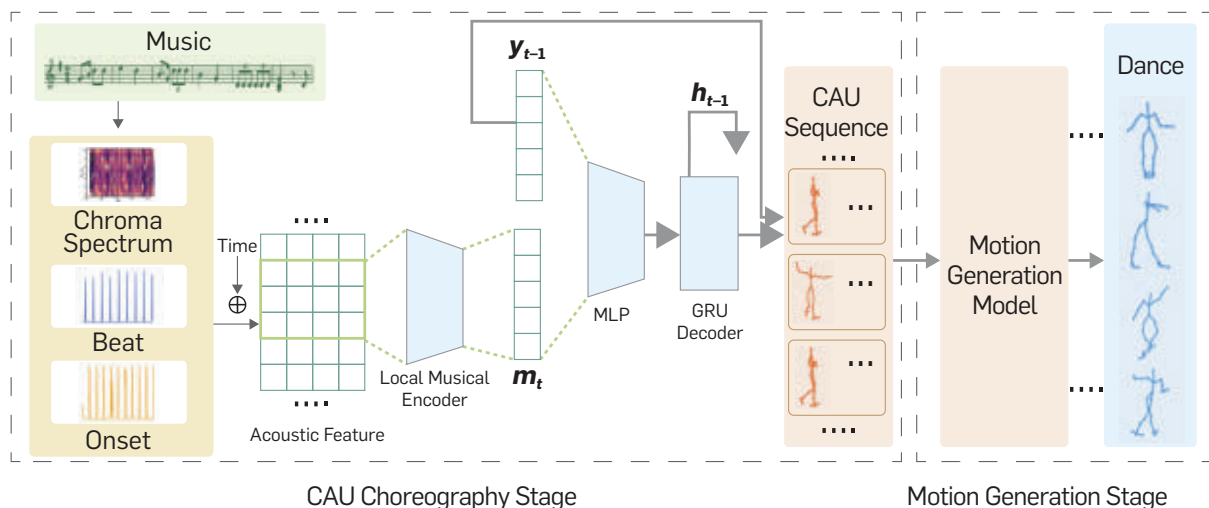
Within a cascade framework, a neural network-based deep discriminant analysis (DDA)<sup>24,26</sup> is suggested to project i-vector to more discriminant embeddings. The direct-embedding framework uses a deep model with more advanced center loss and A-softmax loss, and focal loss is also explored.<sup>25</sup> Moreover, traditional i-vector and neural embeddings are combined with neural network-based DDA to achieve another improvement. Furthermore, AISpeech proposes the use of deep generative models—for example, generative adversarial network (GAN) and variational autoencoder (VAE) models—to perform data augmentation directly on speaker embeddings, which would be used for robust probabilistic linear discriminant analysis (PLDA) training and to improve system accuracy.<sup>2,34</sup> With these newly proposed approaches, the speaker recognition system can significantly improve system robustness and accuracy under noisy and complex conditions.<sup>3</sup>

### Robust TTS

To build robust and highly efficient TTS systems, research on both end-to-end network structures and neural vocoders was conducted. JD proposed an end-to-end speech synthesis framework—duration informed auto-regressive network (DIAN)<sup>19</sup>—which removes the attention mechanism with the help of a separate duration model. This eliminates common skipping and repeating issues. Efficient WaveGlow (EWG), a flow-based neural vocoder, was proposed in Song et al.<sup>18</sup> Compared with the baseline WaveGlow, EWG can reduce inference time cost by more than half, without any obvious reduction in speech quality. To study mixed lingual TTS systems, we look into speaker embedding and phoneme embedding, and study the choice of data for model training in Xue et al.<sup>30</sup> As shown in Figure 5, cross-utterance (CU) context vectors are used to improve the prosody generation for sentences in a paragraph in end-to-end fashion.<sup>28</sup>

Sogou also proposed an end-to-end TTS framework—Sogou-StyleTTS (see Figure 6)—to synthesize highly expressive voice.<sup>12</sup> For front-end text analysis, a cascaded, multitask BERT-LSTM model is adopted. And the acoustic model is improved over FastSpeech,<sup>14</sup> which is composed of a multilayer transformer encoder-decoder and a duration model. Hierarchical VAE is used to extract

Figure 7. The pipeline of the ChoreoNet.



**VDCNN features novel filter designs, pooling operations, input feature map selection, and padding strategies, all of which lead to more accurate and robust ASR performance.**

prosodic information unsupervised to decouple timbre and rhythm, which are considered as style, and a rhythm decoder, to predict the above prosody information. Using this structure, any timbre and rhythm can be combined to achieve style control and introduce GAN to further improve the sound quality, which brings the distribution of acoustic features closer to real voice. Finally, multiband MelGAN architecture<sup>33</sup> is proposed to invert the Mel spectrogram feature representation into waveform samples. Based on StyleTTS, a text-driven, digital-human generation system is proposed to realize a realistic digital human: a multi-modality, generative technology to model the digital human's voice, expressions, lips, and features jointly.

To generate more realistic facial expressions and lip movements, both face reconstruction and generative models are used to map from text to video frames. Moreover, to generate more expressive actions (Figure 7), Sogou cooperated with Tsinghua Tiangong Laboratory to carry out some exploratory work, such as creating digital-human music. ChoreoNet,<sup>35</sup> a two-stage music-to-dance synthesis framework, imitates human choreography procedures. The framework first devises a CAU prediction model to learn the mapping relationship between music and CAU sequences. Afterward, a spatial-temporal inpainting model is devised to convert the CAU sequence into continuous dance motions.

**Network Compression**

Faced with a need to deploy deep

learning methods on edge devices, model compression without accuracy degradation has become a core challenge. Neural network language models (NNLM) have proven to be fundamental components for speech recognition and natural language processing in the deep learning era. Effective NNLM compression approaches that are independent of neural network structures are therefore of great interest. However, most compression approaches usually achieve a high compression ratio at the cost of significant performance loss. AISpeech proposes two advanced, structured-quantization techniques, namely product quantization<sup>16</sup> and soft binarization,<sup>36</sup> to enable the realization of a very high NNLM compression ratio compared to uncompressed models—70–100 without performance loss.<sup>37</sup> The diagram of product quantization for NNLM compression is shown in Figure 8.

**Conclusion**

These research outcomes have been widely used in many areas, including customer service, robotics, and smart home devices. For example, as shown in Figure 9, Xiaoice, originally developed at Microsoft in Beijing, now at XiaoBing.ai, is uniquely designed as an artificial intelligence companion with an emotional connection to satisfy the human need for communication, affection, and social belonging.<sup>17,38</sup> These techniques have successfully driven efficient, sustainable, and stable development, and aim to improve the future of the whole society. 

**Figure 8. Diagram of product quantization for NNLM compression.**

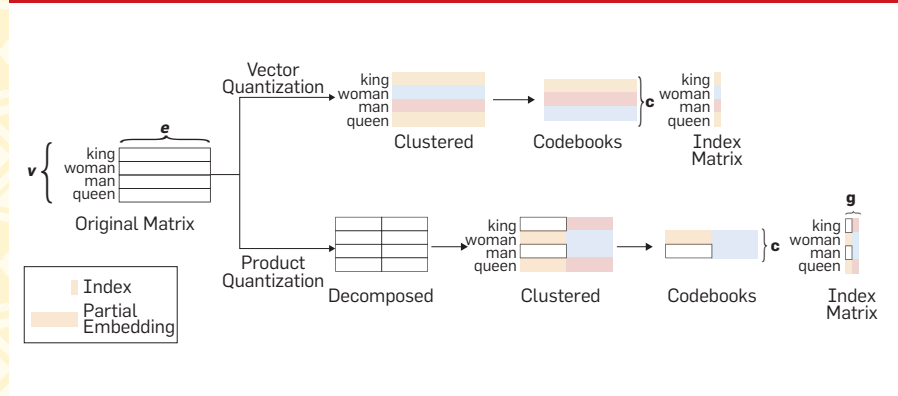
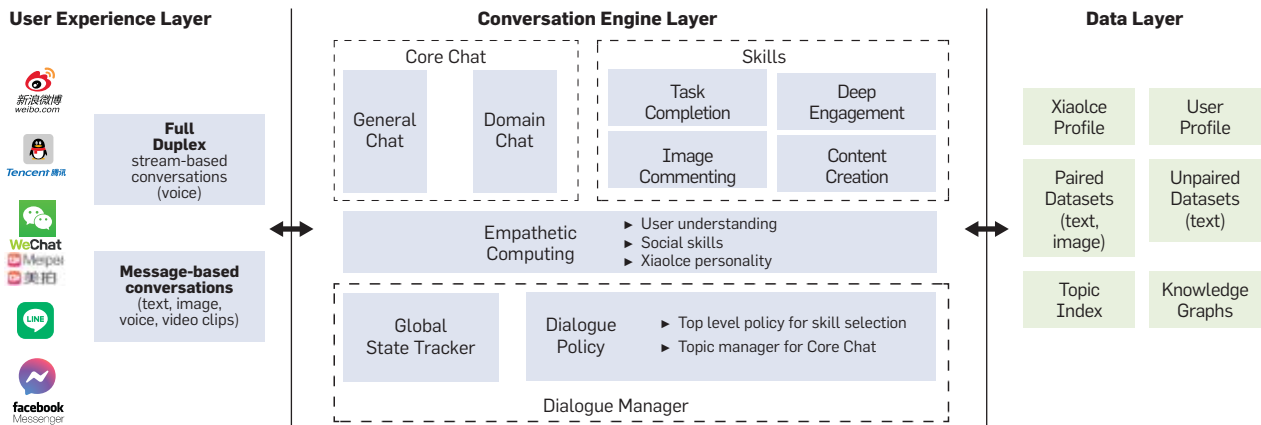




Figure 9. Xiaoice system architecture.



## References

- Bi, M., Qian, Y., and Yu, K. Very deep convolutional neural networks for LVCSR. In *Proceedings of the 16th Annual Conf. Intern. Speech Communication Assoc.*, 2015.
- Chen, Z., Wang, S., and Qian, Y. Adversarial domain adaptation for speaker verification using partially shared network. In *Proceedings of Interspeech 2020*, 3017–3021.
- Chen, Z., Wang, S., Qian, Y., and Yu, K. Channel invariant speaker embedding learning with joint multi-task and adversarial training. In *Proceedings of the IEEE 2020 Intern. Conf. Acoustics, Speech and Signal Processing*, 6574–6578.
- Chung, J., Senior, A., Vinyals, O., and Zisserman, A. Lip reading sentences in the wild. In *Proceedings of the 2017 IEEE Conf. Computer Vision and Pattern Recognition*, 3444–3453.
- Du, J., Tu, Y., Dai, L., and Lee, C. A regression approach to single-channel speech separation via high-resolution deep neural networks. *IEEE/ACM Trans. Audio, Speech, and Language Processing* 24, 8 (2016), 1424–1437.
- Fan, R., Zhou, P., Chen, W., Jia, J., and Liu, G. An online attention-based model for speech recognition. In *Proceedings of Interspeech 2019*, 4390–4394.
- Gao, Z., Zhang, S., Lei, M., and McLoughlin, I. SAN-M: Memory equipped self-attention for end-to-end speech recognition. In *Proceedings of Interspeech 2020*, 6–10.
- Higuchi, Y., Watanabe, S., Chen, N., Ogawa, T., and Kobayashi, T. Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict. In *Proceedings of Interspeech 2020*, 3655–3659.
- Hu, Y. et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. (2020); arXiv:2008.00264.
- Kothapally, V., Xia, W., Ghorbani, S., Hansen, J., Xue, W., and Huang, J. SkipConvNet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping. (2020); arXiv:2007.09131.
- Li, J. et al. Densely connected multi-stage model with channel wise sub-band feature for real-time speech enhancement. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*.
- Meng, F. et al. The Sogou system for Blizzard Challenge. In *Proceedings of 2020 Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 49–53.
- Politis, A., Adavanne, S., and Virtanen, T. Sound event localization and detection task. *2020 DCASE Challenge*; <http://dcase.community/challenge2020/task-sound-event-localization-and-detection-results>
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T. FastSpeech: Fast, robust, and controllable text to speech. *NIPS* (2019), 3165–3174.
- Ryant, N., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. The third DIHARD Speech Diarization Challenge; [https://sat.nist.gov/dihard3#tab\\_leaderboard](https://sat.nist.gov/dihard3#tab_leaderboard)
- Shi, K. and Yu, K. Structured word embedding for low memory neural network language model. In *Proceedings of Interspeech 2018*, 1254–1258.
- Shum, H., He, X., and Li, D. From Eliza to XiaoIce: Challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.
- Song, W., Xu, G., Zhang, Z., Zhang, C., He, X., and Zhou, B. Efficient WaveGlow: An improved WaveGlow vocoder with enhanced speed. In *Proceedings of Interspeech 2020*, 225–229.
- Song, W., Yuan, X., Zhang, Z., Zhang, C., Wu, Y., He, X., and Zhou, B. Dian: Duration informed auto-regressive network for voice cloning. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*.
- Sun, L., Du, J., Gao, T., Fang, Y., Ma, F., and Lee, C. A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the CHiME-5 Challenge. *IEEE J. Selected Topics in Signal Processing* 13, 4 (2019), 827–840.
- Tan, T., Qian, Y., Hu, H., Zhou, Y., Ding, W., and Yu, K. Adaptive very deep convolutional residual network for noise robust speech recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing* 26, 8 (2018), 1393–1405.
- Tong, Y. et al. The JD AI speaker verification system for the FFSVC 2020 Challenge. In *Proceedings of Interspeech 2020*, 3476–3480.
- Vaswani, A. et al. Attention is all you need. (2017); arXiv:1706.03762.
- Wang, S., Huang, Z., Qian, Y., and Yu, K. Discriminative neural embedding learning for short-duration text-independent speaker verification. *IEEE/ACM Trans. Audio, Speech, and Language Processing* 27, 11 (2019), 1686–1696.
- Wang, S., Qian, Y., and Yu, K. Focal KL-divergence based dilated convolutional neural networks for co-channel speaker identification. In *Proceedings of 2018 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*, 5339–5343.
- Wang, S., Yang, Y., Wu, Z., Qian, Y., and Yu, K. Data augmentation using deep generative models for embedding based speaker recognition. *IEEE/ACM Trans. Audio, Speech, and Language Processing* 28 (2020), 2598–2609.
- Watanabe, S., Mandel, M., Barker, J., and Vincent, E. The 6th CHiME Speech Separation and Recognition Challenge (2020); <https://chimechallenge.github.io/chime6/results.html>
- Xu, G., Song, W., Zhang, Z., Zhang, C., He, X., and Zhou, B. Improving prosody modelling with cross-utterance BERT embeddings for end-to-end speech synthesis. In *Proceedings of 2021 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*.
- Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. on Audio, Speech, and Language Processing* 23, 1 (2014), 7–19.
- Xue, L., Song, W., Xu, G., Xie, L., and Wu, Z. Building a mixed-lingual neural TTS system with only monolingual data. In *Proceedings of Interspeech 2019* 2060–2064.
- Xue, W., Qian, G., Zhang, C., Ding, G., He, X., and Zhou, B. Neural kalman filtering for speech enhancement. 2020; arXiv:2007.13962.
- Xue, W., Tong, Y., Zhang, C., Ding, G., He, X., and Zhou, B. Sound event localization and detection based on multiple DOA beamforming and multi-task learning. In *Proceedings of Interspeech 2020*, 5091–5095.
- Yang, G., Yang, S., Liu, K., Fang, P., Chen, W., and Xie, L. Multiband MelGAN: Faster waveform generation for high-quality text-to-speech. In *Proceedings of the 2021 IEEE Spoken Language Technology Workshop*, 492–498.
- Yang, Y., Wang, S., Gong, X., Qian, Y., and Yu, K. Text adaptation for speaker verification with speaker-text factorized embeddings. In *Proceedings of the 2020 IEEE Intern. Conf. Acoustics, Speech and Signal Processing*, 6454–6458.
- Ye, Z., Wu, H., Jia, J., Bu, Y., Chen, W., Meng, F., and Wang, Y. ChoreoNet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM Intern. Conf. Multimedia (2020)*, 744–752.
- Yu, K., Ma, R., Shi, K., and Liu, Q. Neural network language model compression with product quantization and soft binarization. *IEEE/ACM Trans. Audio, Speech, and Language Processing* 28 (2020), 2438–2449.
- Zhao, Z., Liu, Y., Chen, L., Liu, Q., Ma, R., and Yu, K. An investigation on different underlying quantization schemes for pre-trained language models. In *Proceedings of 2020 CCF International Conf. Natural Language Processing and Chinese Computing*. Springer, 359–371.
- Zhou, L., Gao, J., Li, D., and Shum, H. The design and implementation of Xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.
- Zhou, P., Fan, R., Chen, W., and Jia, J. Improving generalization of transformer for speech recognition with parallel schedule sampling and relative positional embedding. (2019); arXiv:1911.00203.
- Zhou, P., Yang, W., Chen, W., Wang, Y., and Jia, J. Modality attention for end-to-end audio-visual speech recognition. In *Proceedings of the 2019 IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, 6565–6569.

**Jia Jia**, Tsinghua University, Beijing, China.

**Wei Chen**, Sogou Corporation, Beijing, China.

**Kai Yu**, Shanghai Jiao Tong University, Shanghai, China.

**Xiaodong He**, JD AI Research, Beijing, China.

**Jun Du**, University of Science and Technology of China, Hefei, China.

**Heung-Yeung Shum**, XiaoBing.ai, Beijing, China.

© 2021 ACM 0001-0782/21/11

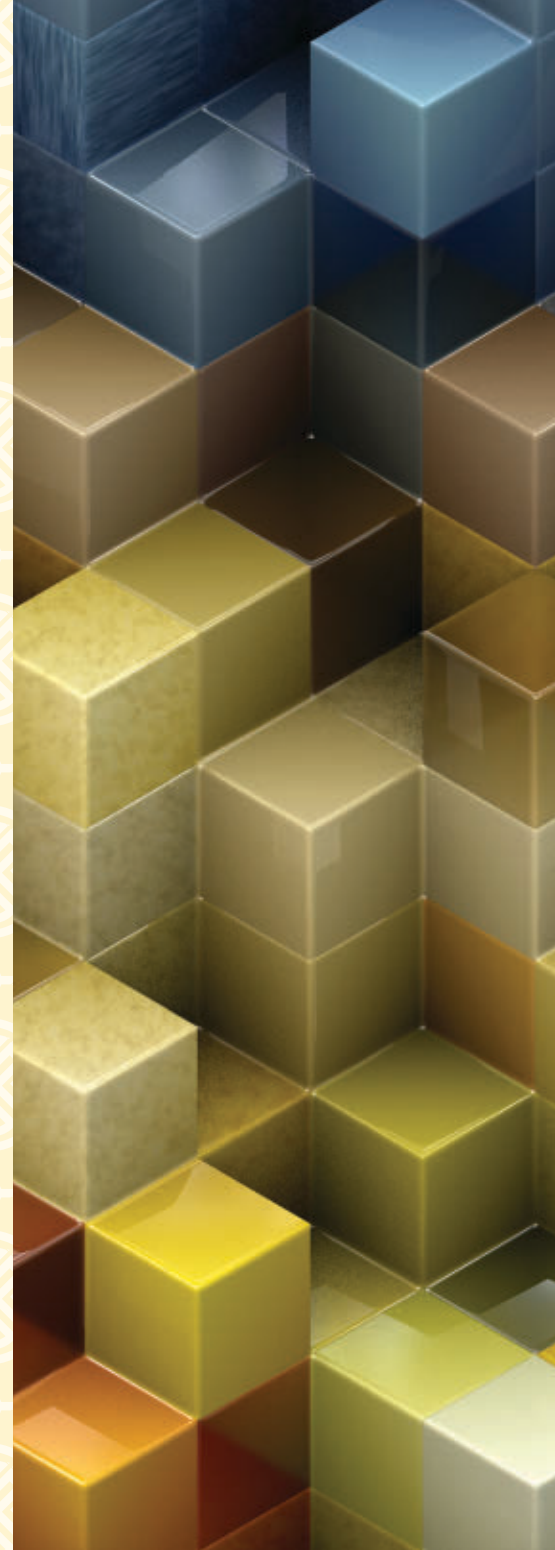
BY LIANG CAI, YI SUN, ZIBIN ZHENG, JIANG XIAO,  
AND WEIWEI QIU

# Blockchain in China

BLOCKCHAIN IS EXPLODING in popularity because of a disruptive fusion of peer-to-peer communication, distributed storage, cryptographic algorithms, and smart contracts. It has become a new paradigm of building trustful distributed systems by providing reliable data service for untrusted parties.

As a trustful distributed system, blockchain has multiple advantages, including immutability, transparency, auditability, and tamper resistance. Any participant can view the complete block and trace each state's transitions. With the security guarantee inherited from the chain-like structure and the consensus mechanism, blockchain records are resistant to malicious forgeries.<sup>17</sup>

With its unique characteristics, blockchain offers strong potential for reestablishing public confidence and enabling reliable information sharing and value transfer. The profound implications of blockchain allow new ways of economic cooperation and have influence over social infrastructure. It is quickly becoming a crucial strategic deployment globally. In the future, blockchain may become a new infrastructure to provide credible and essential data services for enterprises and the public.



## China's Perspectives on Blockchain

China underscores the critical role of blockchain technology in the new round of technological innovation and industrial transformation. However, different from other countries, China has its perspectives on the rise of the worldwide blockchain landscape: embracing the blockchain technology while resisting illegal financial activities related to coin-offering fundraising and trading or "virtual currencies."<sup>3,18</sup>

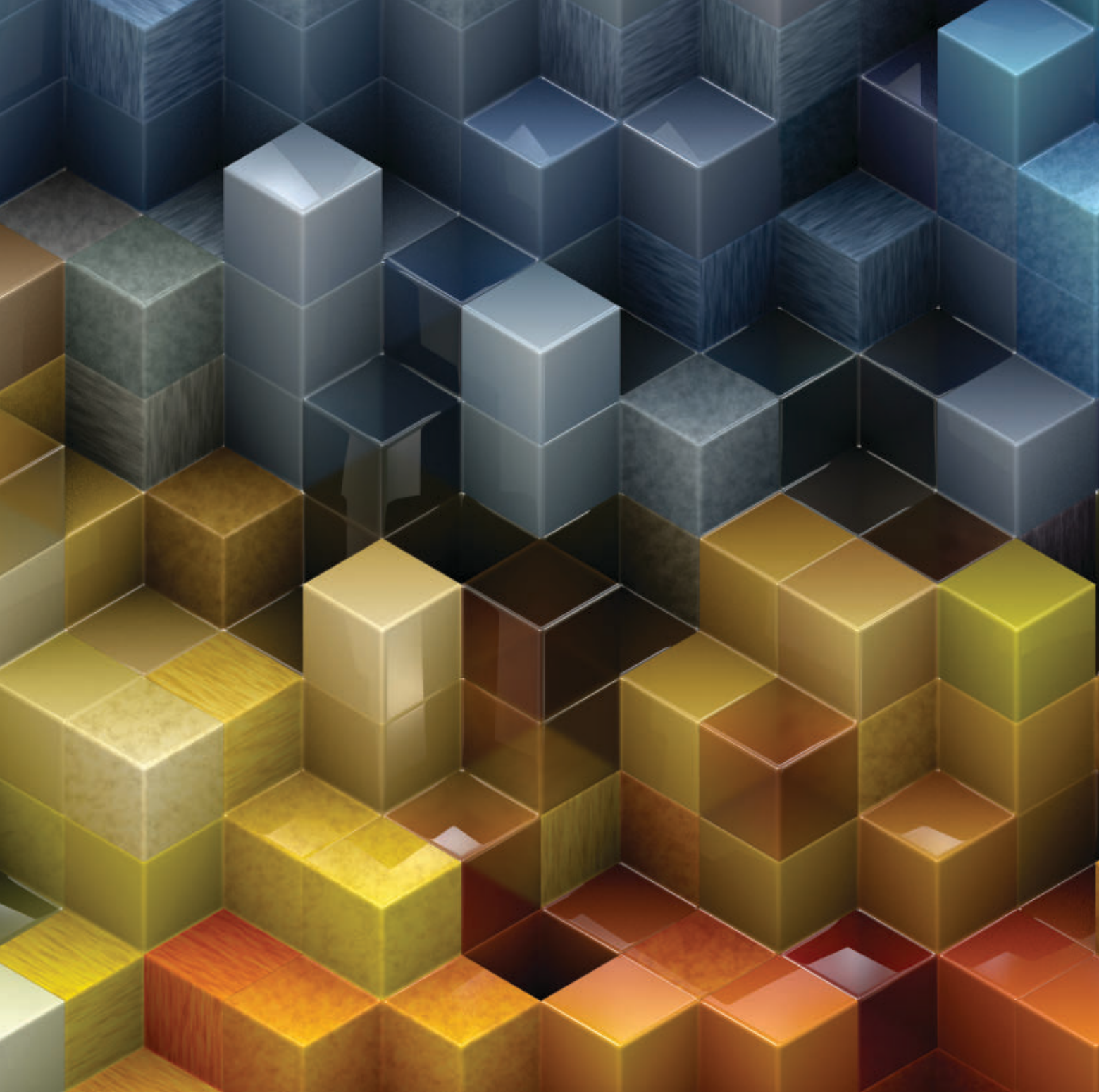


IMAGE BY ALAA ABUMADI

The People's Bank of China (PBoC) announced "Public Notice of the PBC, CAC, MIIT, SAIC, CBRC, CSRC, and CIRC on Preventing Risks of Fundraising through Coin Offering" in September 2017.<sup>18</sup> However, such regulation on cryptocurrency did not bring about a negative influence on blockchain technology. Two weeks after the announcement was published, China's Ministry of Industry and Information Technology launched the Trusted Blockchain Open Lab,

which aims to promote the exploration of blockchain technology without becoming involved in issuing cryptocurrencies, or the exchanges that trade them.<sup>3</sup>

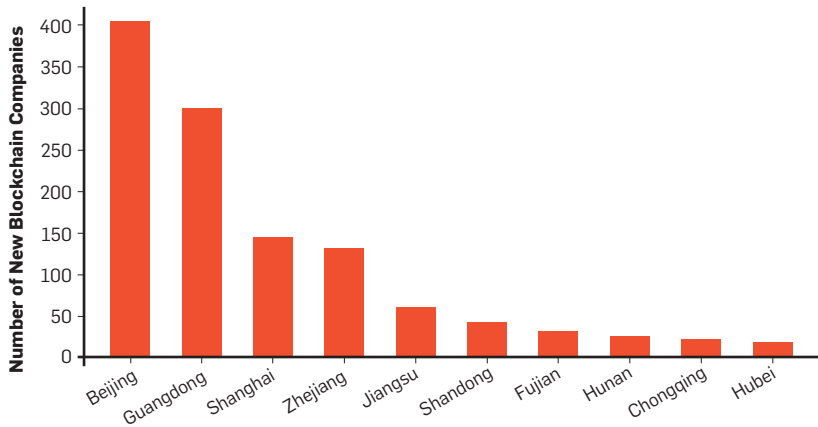
### **The Overall Development Status of Blockchain Technology in China**

After several years of development, China has laid a solid foundation in blockchain technology with the following characteristics.

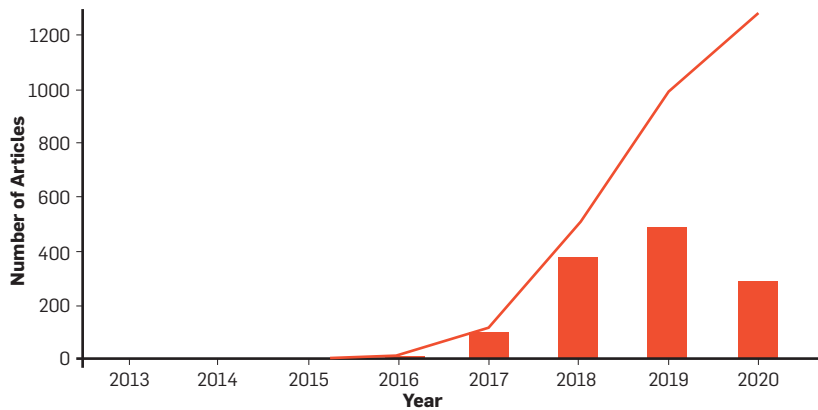
#### **Increasing the R&D capabilities**

**of core technologies.** By the first half of 2020, China had 1,309 companies providing blockchain services<sup>19</sup> and more than 80 blockchain research institutions. Figure 1 indicates most of these companies are concentrated in economically developed areas such as Beijing, Guangdong, and Shanghai,<sup>20</sup> which have the most scientific research talents and R&D resources in China. And in the academic field, as shown in Figure 2, the number of Chinese blockchain-related papers has in-

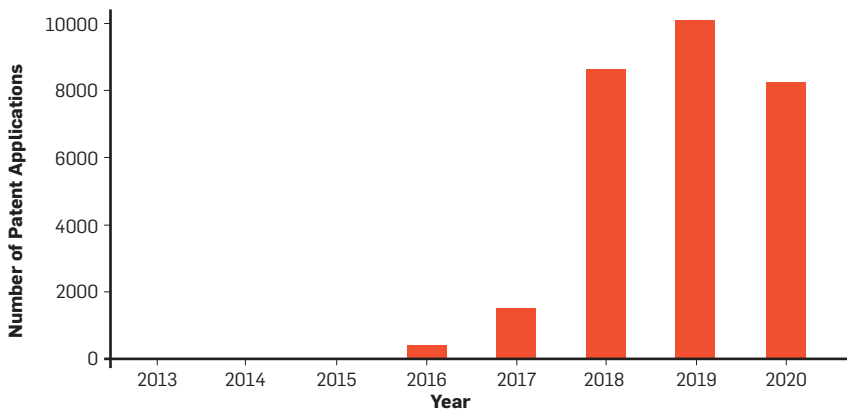
**Figure 1. Top 10 Chinese provinces with the number of blockchain companies.<sup>20</sup>**



**Figure 2. Total amount of blockchain papers of China 2013–Oct. 2020.<sup>13</sup> The bar graph indicates the number of new papers added each year, and the connection indicates the cumulative increase.**



**Figure 3. The number of patent applications in China 2013–2020. (Due to the lag of patent disclosure, the actual data may be more than indicated here).<sup>24,25</sup>**



creased to 1,189 as of October 2020.<sup>13,20</sup> Academic research on blockchain has expanded from digital currency to finance, commerce, people’s livelihoods, government affairs, and other fields, consistent with the domestic blockchain industry’s development trend. Due to the emphasis on core technology, many blockchain solution providers have emerged in China.

**Decoupling architecture for better performance.** Chinese research institutions and enterprises are working on decoupling the blockchain architecture to improve performance, including decoupling consensus and validation, decoupling execution and state, decoupling transactions, and so forth. This also allows different blockchain R&D institutions to participate in designing different layers and reducing corporate R&D costs. Besides, the trend in decoupling architecture promotes China to the establishment of a unified standardized system, realizing the connectivity of different layers.

**Integration with other emerging innovative technologies.** To date, various Chinese enterprises focus on promoting the deep integration of blockchain and other emerging innovative technologies, such as cloud computing,<sup>21</sup> IoT, AI,<sup>22</sup> and 5G,<sup>23</sup> into key industries. These rich function combinations can solve the core issues in the traditional industry and further improve production efficiency. With the attempt to integrate blockchain with emerging technologies, China is seeking new forms of infrastructure and creating new economic growth points.

Steadily growing in the number of patents. Chinese companies are accelerating their patent portfolio in blockchain. The number of patent applications in China ranks first in the world. Figure 3 shows that by the end of 2020, the total number of patent applications in China has exceeded 30,000.<sup>24</sup> More than 4,100 companies have contributed to those patent applications, which occupies a significant proportion of global blockchain patent applications.<sup>25</sup>

**Overview of the Blockchain Industry in China**

The Chinese central government points out it is necessary to build a

blockchain industrial ecosystem. Efforts should be made to promote the deep integration of blockchain with the real economy.<sup>26</sup> Major industrial applications are focused on this target. The main stakeholders include authorities, industrial application companies, technical service providers, and research institutions.

**Authorities outline the industrial policy on regulatory systems and application scenarios.** In 2020, China's national ministries and commissions, provincial governments, and provincial capital cities issued 217 policies, regulations, and program documents related to blockchain technology.<sup>19</sup> The Supreme People's Court has solicited opinions from the public on the review rules and the reinforcement identification of blockchain evidence, as well as the review of the authenticity of data before on-chain.<sup>28</sup> Local governments have paid close attention to blockchain technology and actively build pilot areas for blockchain applications.

**Blockchain technology companies have increased efforts to address technical supports for industry development.** The blockchain industry is proliferating. Since 2019, 57% of the 1,000 enterprises involved in the blockchain business are start-ups; only 23% of the 1,000 are traditional IT companies that establish blockchain sectors.<sup>19</sup> The technology companies represented by Qulian Technology provide the underlying technical support and application construction program support for the blockchain industry. Internet giants including Baidu, Alibaba, Tencent, and JD.com also actively develop blockchain technology and expand application layout.

**Financial, energy, and other industries deepen the application and integration of blockchain technology.** Finance is the first field that has widely applied blockchain technology since 2015. In 2020, the People's Bank of China has released the first standard specification for financial blockchain.<sup>27</sup> The applications in trade finance, supply chain finance, and other scenarios have gradually entered the depth stage. The Forbes Global Blockchain Top 50 2021 listed leading platforms employing distrib-


uted ledger technology, including China Construction Bank's BC Trade 2.0 and Industrial and Commercial Bank of China's 30 blockchain applications.<sup>1</sup> Meanwhile, in the energy field, State Grid proposed the "one main and two sides" blockchain service,<sup>16</sup> to serve renewable energy markets and promote marketization of distributed electricity generation. The in-depth applications, in turn, put forward new requirements for blockchain technology.

#### **Typical Cases of Blockchain Applications in China**


The blockchain applications in China are mainly concentrated on scenarios such as finance, government service, smart cities, and so on.

**The trade finance blockchain platform.** It was initially launched by the Digital Currency Institute of the People's Bank of China in September 2018 to serve SMEs from the financial, taxation, and regulatory perspectives.<sup>29</sup> Five business applications and applications of hundreds of branches of 50 commercial banks have been running on this platform currently. In November 2020, the Trade Finance Blockchain Platform and the HKMA's eTradeConnect Platform completed interconnection with each other, which was the first case of a cross-border collaboration across multiple trade finance blockchain platforms. This collaboration significantly impacts trade finance by realizing cross-border trade finance transactions through the data exchange between two platforms for the first time.

**The cross-provincial housing provident fund platform.** The Housing Provident Fund (HPF) is the largest public housing program in China. Cross-provincial Housing Provident Fund Platform is powered by Hyperchain,<sup>30</sup> whose validating peers consist of Branches of China Construction Bank and the Ministry of Housing and Urban-Rural Development (MOHURD). The platform supports residents in real time withdrawing their "Housing Provident Fund" in 303 House Fund Management Centers across China without any manual review. The platform committed over 480 million transactions and



**Chinese professionals are making efforts to strengthen basic research.**



## Chinese blockchain technologies still have a way to go before they can be sufficient to be applied in different types of scenarios.

ran 19.98 million mortgage accounts for the HPF program.<sup>31</sup> The core advantage of employing blockchain in this scenario is that it enhanced the efficiency of cross-functional collaboration by establishing interdepartmental linkage. Instead of exchanging data from different systems, the blockchain platform supports the entities' updating and utilizing data in one system. Ensured by the blockchain, posted data must be immutable so the data collected from the platform by individuals and entities could be trustworthy for data users, such as the State Taxation Administration.

### **Blockchain core system and applications of Xiong'an New Area.**

Xiong'an New Area is a state-level new area in the Baoding area of Hebei, China. It has been officially regarded as "a strategy that will have lasting importance for the millennium to come."<sup>32</sup> Since 2017, Xiong'an New Area has kept exploring the application of blockchain technology. Xiong'an adopts the blockchain applications of non-tax bills, land transfer and management, judicial deposit, supply chain finance, and so on. Xiong'an launched a city-level blockchain core system for applications in 2020. For example, ICBC Information and Technology Co Ltd has teamed up with the government to use blockchain for compensation payments. This system helps reduce the possible intermediate links and avoids diversion or misappropriation of funds. By the end of 2020, more than 160 construction projects and 40,000 migrant workers were paid via blockchain systems.<sup>33</sup>

### **The Core Technologies for the Blockchain Development in China**

Aggregating the characteristics of the cases introduced here, a practical blockchain platform, which would serve the Chinese economy, should: bear a large number of accounts and transaction throughputs; provide efficient and reliable storage model and network architecture; optimize contract virtual machine; preserve network security and privacy; and support connectivity among het-

erogeneous blockchains. Chinese professionals are making efforts to strengthen basic research and so far, the following innovations were achieved.

**Scalable and high-performance consensus protocols.** Consensus protocols based on permissioned blockchains can be classified into two categories: semi-synchronous and asynchronous protocols. Enterprise-level permissioned blockchains mainly adopt semi-synchronous protocols. Chinese researchers have made some advancements, such as Monoxide,<sup>12</sup> Dumbo-BFT,<sup>4</sup> Conflux,<sup>10</sup> and Pyramid.<sup>7</sup> The most optimized algorithms have linear message complexity and fewer commit phases.

**Cooperative storage model.** To bear massive data storage, a cooperative storage model (for example, Jidar,<sup>2</sup> BFT-Store,<sup>11</sup> ElasticChain,<sup>8</sup> and CUB<sup>14</sup>) can be employed to reduce the storage overhead and improve scalability. Further efforts are made to support SQL syntax and data on a blockchain platform.

**Efficient and reliable network architecture.** To boost the blockchain network's scalability, professionals draw support from the relay network to innovate a multilayer network architecture suitable to the blockchain network (for example, Shrec<sup>5</sup> and BlockP2P<sup>6</sup>). Parallel blockchain structures consisting of a system blockchain and several application blockchains have also been proposed.

**Network security and privacy preservation.** It is essential to guarantee the access control of the blockchain network and privacy-preserving data utility.<sup>15</sup> MSP mechanism is widely used in permissioned blockchains to verify identities and manage the membership of a blockchain network. Moreover, advanced technology such as Zero-Knowledge Proof and Multi-Parties Computing have also been equipped to achieve privacy-preserving data utility.

**Connectivity among heterogeneous blockchains.** As infrastructure, the network should establish value and trust connectivity among those heterogeneous blockchains.<sup>9</sup> Relay-chain architecture is the mainstream blockchain interoperability solution. Qulian Technology Co., Ltd, for

instance, provides open source and self-adaptive inter-blockchain communication solutions based on the relay-chain architecture for heterogeneous blockchains.

**Smart contract and virtual machine optimization.** In addition to Solidity, to support the industrial applications, the blockchain platforms provide various contract virtual machines for popular languages, such as Java and GoLang. For example, TEE can launch smart contracts to provide secure and trustworthy execution of private transactions. Meanwhile, smart contract security is also a research hotspot, whose target is to reduce the risks of vulnerabilities, attacks, and problematic constructs of smart contracts.

## Conclusion

As blockchain was included in the latest 14<sup>th</sup> five-year plan<sup>34</sup> of China, the Chinese central government is continuously promoting the blockchain industry's development, by, for example, releasing supportive standards and regulatory policies, and offering talent training. This provides new opportunities for the development of blockchain.

**Continuously resolving fundamental technical problems.** Chinese blockchain technologies still have a way to go before they can be sufficient to be applied in different types of scenarios. Key enterprises and research institutes must strengthen collaboration to continuously make breakthroughs in core technologies, such as consensus protocols, distributed storage, and P2P network, to further improve the scalability and interoperability of blockchain.

**Gradually forming technical standards and systems.** At present, the blockchain industry in China has a good foundation for development. The national policy settings will give a solid impetus for blockchain development and the rapid formation of technical standards and systems. By 2020, China has issued three blockchain industry standards, five provincial and local standards, and 34 group standards. The Chinese government is drafting three blockchain national standards to guide application practices.<sup>35</sup>

**Barbaric growth has gone to an**

**end.** With the continuous advancement of blockchain supervision and standardization, markets will pay more attention to core technologies' capabilities. The competitive blockchain enterprises will gradually emerge, which marks the latest stage of the industry development. This encourages companies to increase their R&D investment, explore the integration between blockchain and traditional economic modes, and consider the role that blockchain can play in new business models and infrastructure in the future.

**The continuous influx of blockchain talents.** In recent years, China paid much attention to cultivating blockchain talents, leading to the emergence of increasing numbers of technical talents. As of 2020, more than 40 Chinese universities have launched blockchain majors or related courses,<sup>36</sup> and various provinces have also issued more than 30 blockchain talent policies to accelerate their blockchain industry development.<sup>37</sup> The competition and opportunities brought by the influx of talents are beneficial for developing the blockchain industry. 

## References

1. del Castillo, M. Blockchain 50. *Forbes Magazine*. (2021); <https://bit.ly/3BpY8aP>
2. Dai, X. et al. Jidar: A jigsaw-like data reduction approach without trust assumptions for bitcoin system. In *Proceedings of the 2019 IEEE 39<sup>th</sup> Intern. Conf. on Distributed Computing Systems*, Dallas, TX, USA.
3. Dong, L. What's the Future of Blockchain in China? *World Economic Forum*. 2018; <https://bit.ly/36Sll7u>
4. Guo, B. et al. Dumbo: Faster asynchronous BFT protocols. In *Proceedings of the 2020 ACM SIGSAC Conf. Computer and Communications Security*, New York, NY, USA.
5. Han, Y. et al. Shrec: Bandwidth-efficient transaction relay in high-throughput blockchain systems. In *Proceedings of the 11<sup>th</sup> ACM Symp. Cloud Computing*, New York, NY, USA, 2020.
6. Hao, W. et al. BlockP2P: Enabling fast blockchain broadcast with scalable peer-to-peer network topology. In *Proceedings of the 2019 Intern. Conf. on Green, Pervasive, and Cloud Computing*, Uberlandia, Brazil.
7. Hong, Z. et al. Pyramid: A layered sharding blockchain system. In *Proceedings of 2021 Intern. Conf. Computer Communications*.
8. Jia, D. et al. ElasticChain: Support very large blockchain by reducing data redundancy. In *Proceedings of the Asia-Pacific Web and Web-Age Information Management Joint Intern. Conf. on Web and Big Data*, Beijing, China, 2018.
9. Jin, H., Dai, X., and Xiao, J. Towards a novel architecture for enabling interoperability amongst multiple blockchains. In *Proceedings of the 2018 IEEE 38<sup>th</sup> Intern. Conf. on Distributed Computing Systems*, Vienna, Austria.
10. Li, C. et al. A decentralized blockchain with high throughput and fast confirmation. In *Proceedings of the 2020 USENIX Annual Technical Conf.*
11. Qi, X. et al. BFT-Store: Storage partition for permissioned blockchain via erasure coding. In *Proceedings of the IEEE 36<sup>th</sup> Intern. Conf. Data Engineering*, Dallas, TX, USA, 2020.
12. Wang, J. and Wang, H. Monoxide: Scale out blockchain

- with asynchronous consensus zones. In *Proceedings of the 16<sup>th</sup> USENIX Conf. on Networked Systems Design and Implementation*, Boston, MA, USA, 2019.
13. Wang, Q., Su, M. and Li, R. Is China the world's blockchain leader? Evidence, evolution and outlook of China's blockchain research. *J. Cleaner Production* 264, 121742, 2020.
14. Xu, Z., Han, S., and Chen, L. CUB, a consensus unit-based storage scheme for blockchain system. In *Proceedings of the IEEE 34<sup>th</sup> Intern. Conf. on Data Engineering*, Paris, France, 2018.
15. Yan, Y. et al. Confidentiality support over financial grade consortium blockchain. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA, 2020.
16. Yang, D. *Exploration and Practice of Energy Blockchain*. China Electric Power Press, 2020.
17. Zheng, Z. et al. Blockchain challenges and opportunities: A survey. *Intern. J. Web and Grid Services* 14, 4 (2018), 352–375.

## Further Reading

18. People's Bank of China. Public Notice of the PBC, CAMIT, SAIC, CBRC, CSRC and CIRC on Preventing Risks of Fundraising through Coin Offering. People's Bank of China Announcement. 2017.
19. China Academy of Information and Communications Technology. Blockchain White Paper. 2020; <https://bit.ly/3zlT9G9>
20. Tsinghua University Internet Industry Research Institute. China's Blockchain Industry Ecological Map Report. 2021; <https://bit.ly/3xVEI55>
21. HUAWEI CLOUD Officially Launches Blockchain Service for Users Around the World. 2018; <https://bit.ly/3zlx5vh>
22. China's Courts Use Data Analytics and Blockchain Evidence Storage on the Way to First AI-Integrated Legal System. 2021; <https://bit.ly/2UzH0Pd>
23. China Telecom to Develop Blockchain Smartphones. 2019; <https://yhoo.it/36P0n9l>
24. Interpretation of Chinese Blockchain Patent Data. 2021; <https://bit.ly/3wXN15g>
25. Trust Blockchain Initiatives. White Paper on Blockchain Innovation and Intellectual Property Development. 2020; <https://bit.ly/36NNdJY>
26. Xinhua. Xi Stresses Development, Application of Blockchain Technology; <https://bit.ly/3wWoM7B>
27. The People's Bank of China. Financial Distributed Ledger Technology Security Specification. 2020; <https://bit.ly/3y2N1CI>
28. The Supreme People's Court. Provisions on Some Issues of Online Handling of Cases by People's Courts, Draft for Comments. 2021; <https://bit.ly/36P3pe3>
29. CGTN's Global Business. PBOC Research Arm: Blockchain Platform Shortens Trade Finance Process. 2020; <https://bit.ly/36YtA1D>
30. Hyperchain; <https://bit.ly/3hXagBI>
31. The Housing Provident Fund Platform. 2019; <https://bit.ly/2W6MQbj>
32. Xinhua. China Publishes Master Plan for Xiong'an New Area. 2018; <https://bit.ly/3BvHF1A>
33. ChinaDaily. Xiong'an Leads in Blockchain Tech Use. 2020; <https://bit.ly/3xVGKbT>
34. Xinhua. The Development Plan of the 14<sup>th</sup> Five-Year Plan National Strategic Emerging Industry. 2020; <https://bit.ly/2TWAHLW>
35. Blue Book of China Blockchain Standards. 2020; <https://bit.ly/3BnCF2l>
36. People's Daily. The First Blockchain Undergraduate Program in China. 2020; <https://bit.ly/2W1gbUL>
37. People Capital. China's Blockchain Policy Status and Trend Analysis. 2019; <https://bit.ly/3zjNfMc>

**Liang Cai** is a professor and executive deputy director of Blockchain Research Institute at Zhejiang University, Hangzhou, China.

**Yi Sun** is a professor at Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

**Zibin Zheng** is a professor at Sun Yat-sen University, Guangzhou, China.

**Jiang Xiao** is an associate professor at Huazhong University of Science and Technology, Wuhan, China.

**Weiwei Qiu** is the chief architect at Hangzhou Qulian Technology Ltd., Hangzhou, China.

© 2021 ACM 0001-0782/21/11

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

## Why we need an IT accident investigation board.

BY POUL-HENNING KAMP

# What Went Wrong?

IN APRIL, 39 postmasters and sub-postmasters were cleared of wrongdoing by a court in the U.K. after being accused and sentenced for various forms of fraud and, in some cases, serving multiyear prison sentences.<sup>a</sup>

In total, around 700 people have been prosecuted based on the “evidence” from a single IT system installed by the U.K. Post Office, and while some of them probably did embezzle money, it looks like the majority did not. They were sentenced based on evidence from an IT system, which ... eh... to be honest, we don’t know what that IT system did, except we know it did it really, really badly.

<sup>a</sup> <https://www.bbc.com/>

Press reports have contained various mumblings and hand-waving about the shortcomings of the IT system, but nobody sat down and documented precisely what went wrong and what can be learned from it so that nobody ever makes a mistake like this again.

Had this been a ship sinking, a train derailling, or a plane crash, one of the U.K.’s official accident investigation boards would have come in and written a report everybody would be allowed to read, explaining what went wrong and how to avoid it ever happening again. But because no ships, trains, or airplanes were involved, there will be no such report.

For well over a decade, I have been arguing that governments should create IT accident investigation boards for the exact same reasons they have done so for ships, railroads, planes, and in many cases, automobiles.

Denmark got its Railroad Accident Investigation Board because too many people were maimed and killed by steam trains, and it has kept the board around because a thousand tons of steel hurtling along at 180km/h, just below a 25kV power line, can do a lot more damage than a steam locomotive with wooden wagons ever could.

The U.K.’s Air Accidents Investigation Branch was created for pretty much the same reasons, but, specifically, because when the airlines investigated themselves, nobody was any the wiser.

Does that sound slightly familiar in any way?

The crucial feature of any accident investigation board is that it focuses only on what went wrong and how to avoid it happening again, and not on whom to blame.

Sometimes the board may find out that somebody failed to do something crucial, did something illogical, or even did something stupid, but that information is published only if it is necessary to prevent the same type of accident from happening again.

As far as I have seen, the information is relayed in impersonal terms





(“The pilot did ...,” “The clerk did not ...”), because it is not important who that person was; what is important is that no other person exacts that consequence again.

There are three kinds of incidents an IT accident investigation board should look into:

- ▶ when an IT system is involved in loss of life, limb, or liberty;
- ▶ when development of an IT system fails spectacularly; and
- ▶ when an IT system leaks personal information.

The first point is a matter of consistency. Two Boeing 737 MAX airplanes crashed because of IT systems, and because those IT systems happened

to be installed in airplanes, we get reports, whereas we get no reports about the U.K. Post Office’s IT problem because its system was bolted into 19-inch racks.

That makes no sense: The human toll caused by both IT accidents is way beyond anything any civilized society can just let pass.

The second point is a matter of sound fiscal policy. Denmark, like all other countries, has an abysmal track record with development of governmental IT systems. Millions, and in some cases billions, in tax money pour into projects that almost invariably run late, over budget, fail to deliver, and so on.

But nobody is being paid to—or given sufficient access to—write a technical report detailing the crucial mistakes and how to avoid and prevent them in future projects. If an IT accident investigation board were to write a report when such a project failed, and if the contracts for all future projects stipulated that recommendations from the board must be followed, then at least taxpayers would not have to pay to repeat the same mistakes.

The third point should barely need mentioning: Personal information is the helium of IT systems—it leaks out of every crack or imperfection faster than seems possible. This is obviously a subclass of “loss of liberty,” but it is

so dominating that it deserves its own category.

While pretty much everybody agrees that something must be done, nobody wants to give an official IT accident investigation board the authority to find out what that “something” should be. Software houses hem and haw about how their trade secrets and intellectual property will be violated. What they really mean to say is they don’t want anybody to stop their gravy train.

Individual developers fear they will be made scapegoats, even though this is precisely *not* what accident investigation boards do. And politicians and management in private companies are nothing if not unified in their desire to avoid accountability for cutting corners and best-case management.

One particularly bogus argument is that it is not possible to write IT accident reports in the first place. I don’t know where that idea comes from, but surely not from reading accident reports. For example:

In 2017, the motor of an airplane exploded over the southern part of the Greenland icecap. Part of the engine landed on the ice while the plane continued to the first suitable airport way up north in Canada.

Nobody got hurt.

Two years later the accident investigation board located and dug up the missing parts a couple of meters under the surface of Greenland’s ice.


If you think that sounds easy, I highly recommend the 69-page report about how they did it.

A year later, the board issued the final report, revealing that a failure mode called “cold dwell/cold creep” had caused the fan blades to disintegrate. That came as a surprise to everybody, because nobody, not even a mad scientist in a secret lab, had ever imagined that as a failure mode for the Ti-6-4 titanium alloy.<sup>b</sup>


So, yes, surely an IT accident investigation board would find it “impossible” to figure out what went wrong with the U.K. Post IT system. Not!

Another bogus argument is that people would refuse to talk and would destroy and hide evidence. This vastly underestimates lawmakers: It is a crime to do that for all other accident

<sup>b</sup> <https://bit.ly/3ijweWw>



**Software houses hem and haw about how their trade secrets and intellectual property will be violated. What they really mean to say is they don’t want anybody to stop their gravy train.**



investigation boards, and even small infractions lead to jail time. And no, it is not “self-incrimination” unless you did something criminal.

Finally, and most perplexing to me, people claim an IT accident investigation board will cost too much money.

Compared to what?

Compared to destroying the lives of almost 700 people with bogus criminal records and years in jail, separated from their family and kids?

Compared to the 100 million euros Denmark spent on a new IT system for the police, a project that never delivered anything? That amount of money could easily have paid for the first 20 years of a Danish IT accident investigation board.

There really are no valid arguments against IT accident investigation boards, and all the bogus arguments proffered are the same ones that people put forth to counter all the other very successful accident investigation boards now in operation.

These boards work. We need one for IT, and we need it now.

*Note: Shortly after this article was written, the U.S. announced the establishment of a new Cybersecurity Safety Review Board, similar to what is described here.<sup>c</sup>*

*“The Executive Order establishes a Cybersecurity Safety Review Board, co-chaired by government and private sector leads, that may convene following a significant cyber incident to analyze what happened and make concrete recommendations for improving cybersecurity. Too often organizations repeat the mistakes of the past and do not learn lessons from significant cyber incidents. When something goes wrong, the Administration and private sector need to ask the hard questions and make the necessary improvements. This board is modeled after the National Transportation Safety Board, which is used after airplane crashes and other incidents.”* ■

<sup>c</sup> <https://www.whitehouse.gov/>

**Poul-Henning Kamp** (phk@FreeBSD.org) spent more than a decade as one of the primary developers of the FreeBSD operating system before creating the Varnish HTTP Cache software, which around one-fifth of all Web-traffic goes through. He is an independent contractor; one of his most recent projects was a super-computer cluster, to stop the stars twinkling in the mirrors of ESO’s new ELT telescope.

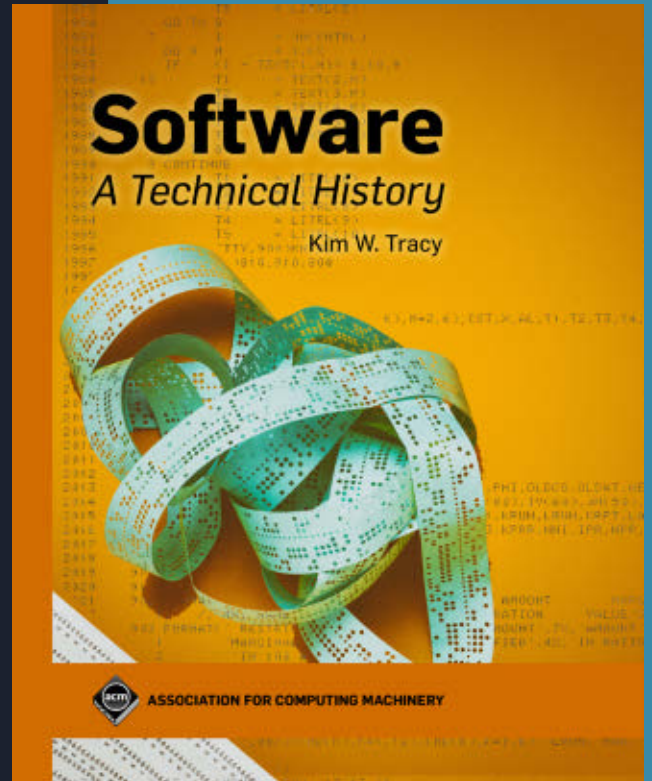


## ACM BOOKS

### Collection II

Software history has a deep impact on current software designers, computer scientists, and technologists. System constraints imposed in the past and the designs that responded to them are often unknown or poorly understood by students and practitioners, yet modern software systems often include “old” software and “historical” programming techniques. This work looks at software history through specific software areas to develop student-consumable practices, design principles, lessons learned, and trends useful in current and future software design. It also exposes key areas that are widely used in modern software, yet infrequently taught in computing programs. Written as a textbook, this book uses specific cases from the past and present to explore the impact of software trends and techniques.

Building on concepts from the history of science and technology, software history examines such areas as fundamentals, operating systems, programming languages, programming environments, networking, and databases. These topics are covered from their earliest beginnings to their modern variants. There are focused case studies on UNIX, APL, SAGE, GNU Emacs, Autoflow, internet protocols, System R, and others. Extensive problems and suggested projects enable readers to deeply delve into the history of software in areas that interest them most.



## Software

### *A Technical History*

**Kim W. Tracy**

ISBN: 978-1-4503-8725-5

DOI: 10.1145/3477339

<http://books.acm.org>

<http://store.morganclaypool.com/acm>

DOI:10.1145/3448247

## Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises.

BY CHRISTOPH GRÖGER

# There Is No AI Without Data

ARTIFICIAL INTELLIGENCE (AI) has evolved from hype to reality over the past few years. Algorithmic advances in machine learning and deep learning, significant increases in computing power and storage, and huge amounts of data generated by digital transformation efforts make AI a game-changer across all industries.<sup>8</sup> AI has the potential to radically improve business processes with, for instance, real-time quality prediction in manufacturing, and to enable new business models,

such as connected car services and self-optimizing machines. Traditional industries, such as manufacturing, machine building, and automotive, are facing a fundamental change: from the production of physical goods to the delivery of AI-enhanced processes and services as part of Industry 4.0.<sup>25</sup> This paper focuses on AI for industrial enterprises with a special emphasis on machine learning and data mining.

Despite the great potential of AI and the large investments in AI technologies undertaken by industrial enterprises, AI has not yet delivered on the promises in industry practice. The core business of industrial enterprises is not yet AI-enhanced. AI solutions instead constitute islands for isolated cases—such as the optimization of selected machines in the factory—with varying success. According to current industry surveys, data issues constitute the main reasons for the insufficient adoption of AI in industrial enterprises.<sup>27,35</sup>

In general, it is nothing new that data preparation and data quality are key for AI and data analytics, as there is no AI without data. This has been an issue since the early days of business intelligence (BI) and data warehousing.<sup>3</sup> However, the manifold data challenges of AI in industrial enterprises go far beyond detecting and repairing dirty data. This article profoundly investi-

### » key insights

- **Despite AI's great potential, the business of industrial enterprises is not yet AI-enhanced. AI is done in an insular fashion, leading to a polyglot and heterogeneous enterprise data landscape that limits the comprehensive application of AI.**
- **Data challenges, such as data management, data democratization, and data governance, constitute the major obstacles to leveraging AI and go far beyond ensuring data quality, comprising diverse aspects such as metadata management, data architecture, and data ownership.**
- **The presented data ecosystem for industrial enterprises addresses these challenges and comprises data producers, data platforms, data consumers, and data roles for AI.**



gates these challenges and rests on our practical real-world experiences with the AI enablement of a large industrial enterprise—a globally active manufacturer. At this, we undertook systematic knowledge sharing and experience exchange with other companies from the industrial sector to present common issues for industrial enterprises beyond an individual case.

As a starting point, we characterize the current state of AI in industrial enterprises, called “insular AI,” and present a practical example from manufacturing. AI is typically done in islands for use case-specific data provisioning and data engineering, leading to a heterogeneous and polyglot enterprise data landscape. This causes various data challenges that limit the comprehensive application of AI.

We particularly investigate challenges to data management, data democratization, and data governance which result from real-world AI projects. We illustrate them with practical examples and systematically elaborate on related aspects, such as metadata management, data architecture, and data ownership. To address the data challenges, we introduce the data ecosystem for industrial enterprises as an overall framework. We detail both IT-

technical and organizational elements of the data ecosystem—for example, data platforms and data roles. Next, we assess how the data ecosystem addresses individual data challenges and paves the way from insular AI to industrialized AI. Then, we highlight the open issues we face in the course of our real-world realization of the data ecosystem and point out future research directions—for instance, the design of an enterprise data marketplace.

**Current State of AI in Industrial Enterprises**

In the following, we define AI and data analytics as key terms and offer an overview of the business of industrial enterprises to concretize the scope of our work. On this basis, we characterize the current state of AI and illustrate it with a practical example.

**Artificial intelligence and data analytics.** Generally, AI constitutes a fuzzy term referring to the ability of a machine to perform cognitive functions.<sup>10</sup> Approaches to AI can be subdivided into deductive—that is, model-driven (such as expert systems)—or inductive—that is, data-driven.<sup>10</sup> In this paper, we focus on data-driven approaches, particularly machine learning and data mining,<sup>17</sup> as they have opened

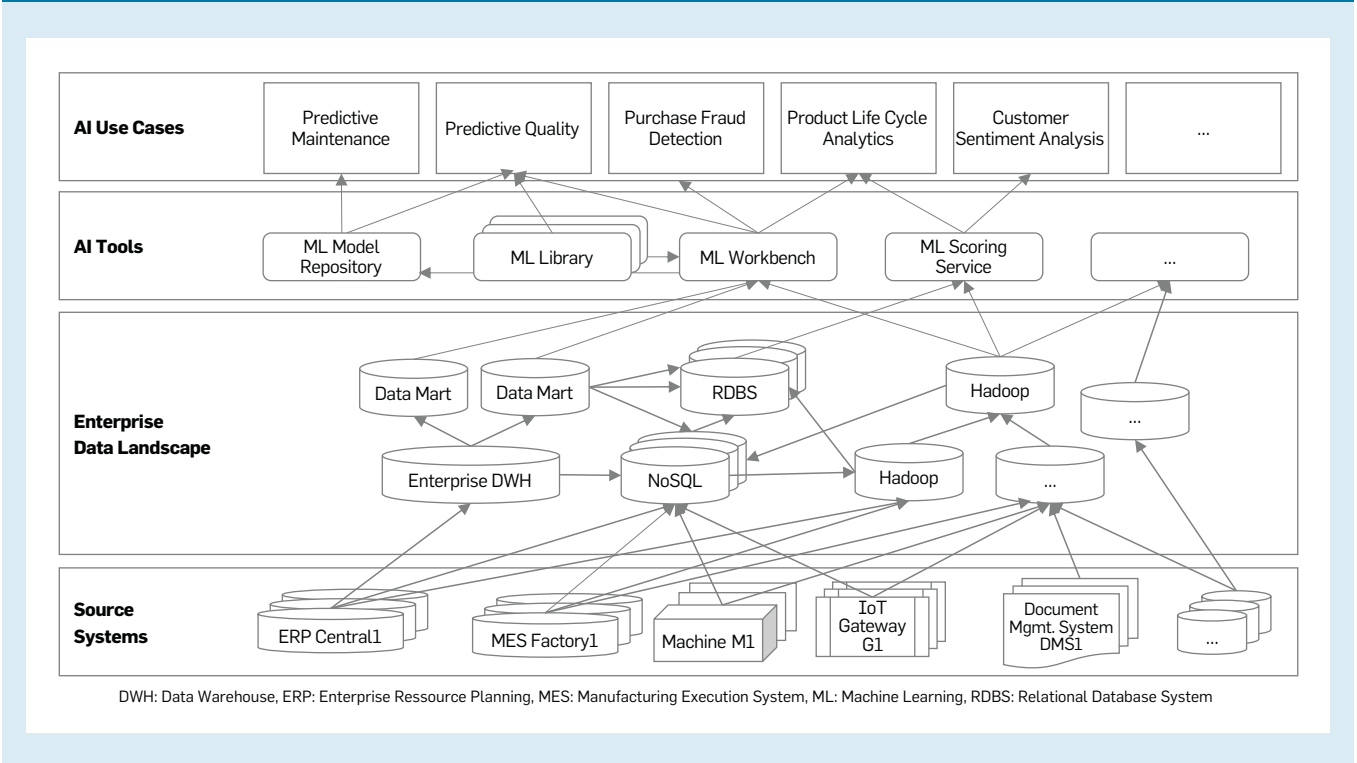
new fields of application for AI in the last years. Moreover, we use data analytics<sup>4</sup> as an umbrella term for all kinds of data-driven analysis, including BI and reporting.

**Business of industrial enterprises.**

The business of industrial enterprises comprises the engineering and manufacturing of physical goods—for instance, heating systems or electrical drives. For this purpose, industrial enterprises typically operate a manufacturing network of various factories organized into business units. The IT landscape of industrial enterprises usually comprises different enterprise IT systems, ranging from enterprise resource planning (ERP) systems over product lifecycle management (PLM) systems to manufacturing execution systems (MES).<sup>24</sup> In Industry 4.0 and Internet of Things (IoT) applications, industrial enterprises push the digitalization of the industrial value chain.<sup>22</sup> The aim is to integrate data across the value chain and exploit it for competitive advantage. Hence, the AI enablement of processes and products is of strategic importance. To this end, industrial enterprises have, in recent years, built data lakes, introduced AI tools, and created data science teams.<sup>15</sup>

**Current state: insular AI.** Figure 1


Figure 1. Current state of AI in industrial enterprises: insular AI with heterogeneous enterprise data landscape.




illustrates the current state of AI in industrial enterprises per the results of our investigations. Organizations have implemented a wide variety of AI use cases across the industrial value chain: from predictive maintenance for IoT-enabled products over predictive quality for manufacturing process optimization to product lifecycle analytics and customer sentiment analysis (see Gröger and Laudon, et al.<sup>15,24</sup> for details on these use cases). The use cases combine data from various source systems, such as ERP systems and MESs, and are typically implemented as isolated solutions for each individual case. That means, AI is performed in “islands” for use case-specific data provisioning and data engineering as well as for use case-specific AI tools and fit-for-purpose machine-learning algorithms. This is what we call “insular AI.”

On one hand, insular AI fosters the flexibility and explorative nature of use-case implementations. On the other hand, it hinders reuse, standardization, efficiency, and enterprise-wide application of AI. The latter is what we call “industrialized AI.” In the rest of this article, we focus on data-related issues of AI because the handling of data plays a central role on the path to industrialized AI. In fact, data handling accounts for around 60% to 80% of the entire AI use case implementation, according to our experiences.

Insular AI leads to a globally distributed, polyglot, and heterogeneous enterprise data landscape (see Figure 1). Structured and unstructured source data for AI use cases is extracted and stored in isolated raw data stores, called data lakes.<sup>13</sup> They are based on individual data storage technologies—for instance, different NoSQL systems, use case-specific data models, and dedicated source-data extracts. These data lakes coexist with the enterprise data warehouse,<sup>23</sup> which contains integrated and structured data from various ERP systems for reporting purposes. The many data-exchange processes in existence cause diverse data redundancies and potential data quality issues. Besides, the disparate data landscape significantly complicates the development of an integrated, enterprise-wide view of business objects—for example, products and processes—and thus hin-



**AI has not yet delivered on the promises in industry practice. The core business of industrial enterprises is not yet AI-enhanced.**



ders cross-process and cross-product AI use cases.

#### **Practical manufacturing example.**

To illustrate the shortcomings of insular AI and underline the need for an overall approach, we take an example from manufacturing. To predict the quality of a specific manufacturing process in a factory, a specialist project team of data scientists and data engineers first identifies relevant source systems, especially several local MESs in the factory as well as a central ERP system. The MESs provide sensor data on quality measurements and the ERP system provides master data. Together with various IT specialists, manufacturing experts, and data owners, the team inspects the data structures of the source systems and develops customized connectors for extracting source data and storing it in the local factory data lake in its raw format.

Data is cleansed, integrated, and pivoted based on a use case-specific data model and various case-specific data pipelines. As a general documentation of the business meaning of individual tables and columns is missing, this is done manually in the project’s internal documents. The team then employs various machine-learning tools to generate an optimal prediction model. Over the course of several iterations, the data model and source-data extracts are adapted to enhance the data basis for machine learning. The final prediction model is then used in the MES on the factory shop floor by calling a machine-learning scoring service.

Overall, the resulting solution constitutes a targeted but isolated AI island with use case-specific data extracts, custom data models, tailored data pipelines, a dedicated factory data lake, and fit-for-purpose machine-learning tools. At this, the solution incorporates a large body of expert knowledge considering manufacturing-process know-how, ERP and MES IT system know-how, use case-specific data engineering, and data science know-how. Yet, missing data management guidelines (such as those for data modeling and metadata management), little transparency on source systems, and the variety of isolated data lakes all hinder reuse, efficiency, and enterprise-wide application of AI. That


is, the same type of use case gets implemented from scratch in different ways across different factories even though it refers to the same type of source systems, the same conceptual data entities, and the same type of manufacturing process. Thus, the same source data—for instance, master data—is extracted multiple times, creating a high load on business-critical source systems, such as ERP systems. Different data models are developed for the same conceptual data entities, such as ‘machine’ and ‘product’. These heterogeneous data models and different data-storage technologies used in individual factory data lakes lead to heterogeneous data pipelines for pivoting the same type of source data, such as MES tables with sensor data. Besides, the business meaning of data and developed data models—that is, metadata—are documented multiple times in project-specific tools, such as data dictionaries or spreadsheets. All in all, this leads to an ocean of AI islands and a heterogeneous enterprise data landscape.

Consequently, to industrialize AI requires a systematic analysis of the underlying data challenges. On this basis, an overall solution integrating technical and organizational aspects can be designed to address the challenges.


### Data Challenges of AI

Based on our practical investigations at the manufacturer, we identified manifold data challenges of AI and systematically clustered them. We aligned these challenges with other companies during systematic knowledge sharing to present common issues for industrial enterprises. Current literature<sup>6,21</sup> and industry surveys<sup>27,35</sup> on AI in industrial enterprises support our findings. Notably, this article goes significantly beyond these related works by analyzing both organizational and technical aspects of the data challenges and by providing detailed industry experiences on the individual challenges.

Generally, ensuring data quality for AI is important—for instance, by detecting and cleansing dirty data. Such data quality issues have already been addressed by a plurality of works and tools.<sup>5,39</sup> Beyond data quality, however, exist further critical data chal-



## According to current industry surveys, data issues constitute the main reasons for the insufficient adoption of AI in industrial enterprises.



lenges—data management, data democratization, and data governance for AI (see Figure 2)—which we focus on in this article. We detail them with special emphasis on data-driven AI—that is, machine learning and data mining. In contrast to classical BI and reporting, machine learning and data mining impose extended data requirements.<sup>6</sup> They favor the use of not only aggregated, structured data but also of high volumes of both structured and unstructured data in its raw format—for example, for machine learning-based optical inspection.<sup>40</sup> This data also needs to be processed not only in periodic batches but also in near real time to provide timely results—for instance, to predict manufacturing quality in real time.<sup>6</sup> Consequently, AI poses new challenges to data management, data democratization, and data governance as detailed in the following.

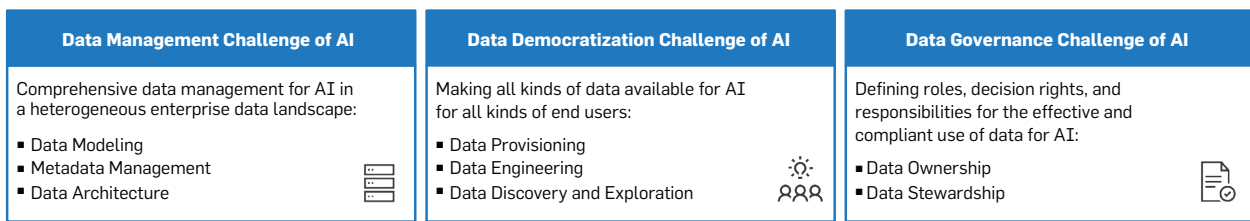
### Data management challenge of AI.

Data management generally comprises all concepts and techniques to process, provision, and control data throughout its life cycle.<sup>18</sup> The data management challenge of AI lies in comprehensively managing data for AI in a heterogeneous and polyglot enterprise data landscape. According to our practical investigations, this particularly refers to data modeling, metadata management, and data architecture for AI.

No common data modeling approaches exist for how to structure and model data on a conceptual and logical level across the data landscape. Frequently, different data-modeling techniques, such as data vault<sup>26</sup> or dimensional modeling,<sup>23</sup> are used for the same kinds of data—for instance, manufacturing sensor data—in the data lakes. Sometimes, even the need for data modeling is neglected with reference to a flexible schema-on-read approach on top of raw data. This significantly complicates data integration, reuse of data, and developed data pipelines across different AI use cases. For instance, pivoting sensor data as input for machine learning is time-consuming and complex. Reusing corresponding data pipelines for different AI use cases significantly depends on common data-modeling techniques and common data models for manufacturing data, in this example.



Figure 2. Data challenges of AI and related aspects.



There is no overall metadata management to maintain metadata across the data landscape. Technical metadata, such as the names of columns and attributes, are mostly stored in the internal data dictionaries of individual storage systems and are not generally accessible. Hence, data lineage and impact analyses are hindered. For instance, in the case of changes in source systems, manually adapting the affected data pipelines across all data lakes without proper lineage metadata is tedious and costly. Moreover, business metadata on the meaning of data—for example, the meaning of KPIs—is often not systematically managed at all. Thus, missing metadata management significantly hampers data usage for AI.

No overarching data architecture structures the data landscape. Missing on one hand is an enterprise data architecture to orchestrate various isolated data lakes. For instance, there is no common zone model<sup>37</sup> across all data lakes, which complicates data integration and exchange. Moreover, the integration of the existing enterprise data warehouse containing valuable key performance indicators (KPIs) for AI use cases is unclear. On the other hand, also lacking is a systematic platform data architecture to design a data lake. Specifically, different data storage technologies are used to realize data lakes. For example, some data lakes are solely based on Hadoop storage technologies, such as HDFS<sup>a</sup> and Hive,<sup>b</sup> while others combine classical relational database systems and NoSQL systems. This leads to non-uniform data-lake architectures across the enterprise data landscape, resulting in high development and maintenance costs.

a <http://hive.apache.org>  
 b <http://hadoop.apache.org>

**Data democratization challenge of AI.** In general, data democratization refers to facilitating the use of data by everyone in an organization.<sup>41</sup> The data democratization challenge of AI lies in making all kinds of data available for AI for all kinds of end users across the entire enterprise. To this end, data provisioning and data engineering as well as data discovery and exploration all play central roles for AI. According to our investigations, these activities are mostly limited to small groups of expert users in practice and thus prevent data democratization for AI as explained in the following.

Data provisioning—that is, technically connecting new source systems to a data lake and extracting selected source data—typically requires dedicated IT projects. To that end, IT experts are concerned with defining technical interfaces and access rights for source systems and developing data extraction jobs in cooperation with source-system owners and data end users. Hence, the central IT department frequently becomes a bottleneck factor for data provisioning in practice. Moreover, there is a huge need for coordination between IT experts, source-system owners, and end users, which leads to time-consuming iterations. These factors significantly slow down and limit data provisioning and thus the use of new data sources for AI.

Data engineering—modeling, integrating, and cleansing of data—is typically done by highly skilled data scientists and data engineers. Due to incomplete metadata on source systems, data engineering requires specialist knowledge of individual source systems and their data structures—for example, on the technical data structures of ERP systems. In addition, mostly complex, script-based frame-

works, such as with Python,<sup>c</sup> are used for data-engineering tasks requiring comprehensive programming knowledge. These factors limit data engineering to small groups of expert users.

This also holds true for data discovery and exploration. Although self-service visualization tools are provided, discovery and exploration of data in data lakes is hampered. Comprehensive metadata on the business meaning and quality of data is missing, preventing easy data usage by non-expert users. For instance, a marketing specialist must identify and contact several different data engineers, who have prepared different kinds of market data, to understand the meaning and interrelations of the data. Besides, compliance approvals for data usage are typically based on specialist inspections of data, such as inspections by legal experts in the case of personal data. These low-automation processes also slow down the use of data for AI.

**Data governance challenge of AI.** Generally, data governance is about creating organizational structures to treat data as an enterprise asset.<sup>1</sup> The data governance challenge of AI refers to defining roles, decision rights, and responsibilities for the economically effective and compliant use of data for AI. According to our practical investigations, organizational structures for data are only rudimentarily implemented in industrial enterprises and mainly focus on master data and personal data. Particularly, structures for data ownership and data stewardship are missing, hampering the application of AI as follows.

There is no uniform data ownership organization across the heterogeneous data landscape. Especially, data own-

c <http://www.python.org>

ership for data extracted and stored in different data lakes is not defined in a common manner. For instance, in many cases, the owner of the data in the data lake remains the same as the data owner of the source system. That is, the integration of data from different source systems stored in the data lake requires approvals by different data owners. Hence, data is not treated as an enterprise asset owned by the company but rather as an asset of an individual business function—for example, the finance department as data owner of finance data. This leads to unclear responsibilities and an unbalanced distribution of risks and benefits when using data for AI.

For example, when manufacturing-process data from an MES is integrated with business-process data from an ERP system to enable predictive maintenance, the respective data owners—for instance, the manufacturing department and the finance department—must agree on and remain liable for a possibly noncompliant use of this data. However, the benefit of a successful use-case implementation, such as lower machine-maintenance costs, is attributed to the engineering department. In other cases, data ownership in the data lake is decoupled from data ownership in source systems to avoid this issue. Yet, this may lead to heterogeneous and overlapping data ownership structures, such as when data ownership is orga-

nized by business function in source systems and by business unit in the data lake. These organizational boundaries significantly hinder the comprehensive use of data for AI.

There is no overall data stewardship organization to establish common data policies, standards, and procedures. Existing data stewardship structures in industrial enterprises mainly focus on various kinds of master data to define—for example, common data quality criteria for master data on customers. Data stewardship for further categories of data is not systematically organized. For example, there are various data models as well as data quality criteria on manufacturing data across different factories and manufacturing processes. Thus, common enterprise-wide policies for manufacturing data are lacking. This significantly increases the efforts and complexity of data engineering for AI use cases.

### Call for a Data Ecosystem for Industrial Enterprises

In light of the above data challenges, we see the need for a holistic framework that covers both technical and organizational aspects to address the data challenges of AI. To this end, we adopt the framework of a data ecosystem. Generally, a data ecosystem represents a socio-technical, self-organizing, loosely coupled system for the sharing of data.<sup>31</sup> A data ecosystem's

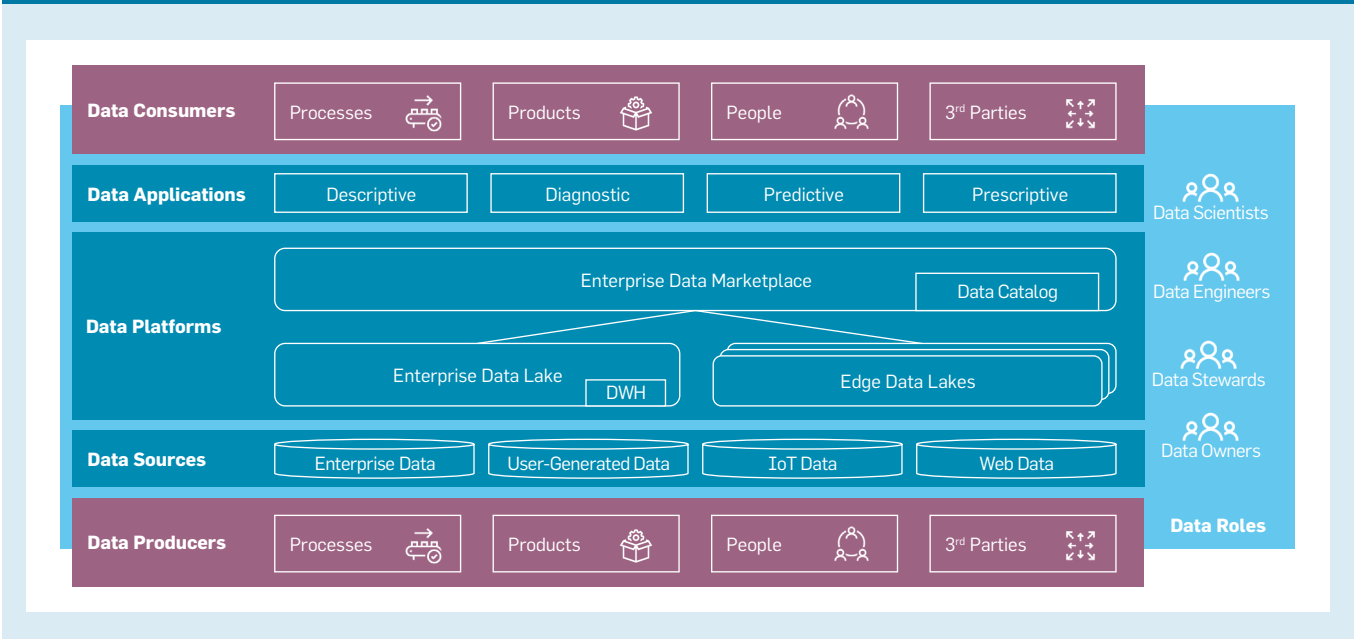
typical elements are data producers, data consumers, and data platforms.<sup>31</sup> However, data ecosystem research is still in its early stages and mainly focused on the sharing of open government data.<sup>33</sup> Therefore, we call for a data ecosystem specifically tailored to industrial enterprises.

Based on our practical experiences with the AI enablement of the manufacturer and knowledge exchange with further industrial companies, we derived core data ecosystem elements for industrial enterprises (see Figure 3). They are described in the following:

**Data producers and data consumers.** Data producers and consumers represent resources or actors generating or consuming data. We generally differentiate four kinds of data producers in an industrial enterprise: Processes refer to all kinds of industrial processes and resources across the value chain—for instance, engineering processes.<sup>24</sup> Products refer to manufactured goods, such as electrical drives or household appliances. People comprise all kinds of human actors, including customers and employees. Third parties comprise actors and resources outside the organizational scope of the enterprise—for example, suppliers.

**Data sources.** Data sources relate to the technical kind and the sources of data generated by data producers. We distinguish between four kinds of data

Figure 3. Core elements of a data ecosystem for industrial enterprises.



sources in an industrial enterprise: Enterprise data refers to all data generated by enterprise IT systems across the industrial value chain, such as PLM and ERP systems.<sup>24</sup> User-generated data refers to data directly generated by human actors, such as social media postings or documents. IoT data refers to all data generated by IoT devices, such as manufacturing machine data or sensor data.<sup>6</sup> Web data refers to all data from the Web, except user-generated data—for instance, linked open data or payment data.

**Data platforms.** Data platforms represent the technical foundation for data processing from all kinds of data sources to make data available for various data applications. The data ecosystem is based on three kinds of data platforms: the enterprise data lake, edge data lakes, and the enterprise data marketplace.

The enterprise data lake constitutes a logically central, enterprise-wide data lake. It combines the original data lake approach<sup>29</sup> with the data warehouse concept.<sup>23</sup> That means, it combines the data lake-like storage and processing of all kinds of raw data with the data warehouse-like analysis of aggregated data. Batch and stream data processing are supported to enable all kinds of analyses on all kinds of data. The enterprise data lake is based on comprehensive guidelines for data modeling and metadata management and enables enterprise-wide reuse of data and data pipelines.

Edge data lakes represent decentralized raw data stores that complement the enterprise data lake. Edge data lakes focus on the realization of data applications based on local data, with little enterprise-wide reuse. They are particularly suited for data processing in globally distributed factories, with selected factories operating their own edge data lake. A typical AI use case for edge data lakes is to predict time-series data produced by a specific manufacturing machine in a single factory of the enterprise.

The enterprise data marketplace constitutes the central pivot point of the data ecosystem. It represents a metadata-based self-service platform that connects data producers with data consumers. The goal is to match supply and demand for data within



**Based on our practical experiences with the AI enablement of the manufacturer and knowledge exchange with further industrial companies, we derived core data ecosystem elements for industrial enterprises.**



the enterprise. However, research on data marketplaces is at an early stage and there are only initial concepts focusing on external enterprise marketplaces for data.<sup>36,38</sup> Hence, we work out essential characteristics of an internal enterprise data marketplace fitting the data ecosystem.

In contrast to the enterprise data lake and edge data lakes, the enterprise data marketplace does not store the actual data. Rather, it is based on a data catalog<sup>37</sup> representing a metadata-based inventory of data. That is, data is represented by metadata and a reference to the actual data. For instance, the data catalog item, “Quality Data for Product P71” might comprise metadata on the related product and a reference to a set of sensor data stored in the enterprise data lake. Data catalog items not only refer to data in the data lakes but also to data in source systems, such as ERP and PLM systems. Besides, metadata from application programming interfaces (APIs) that expose data are also fused in the data catalog. Hence, the marketplace in combination with the data catalog provides a metadata-based overview of all data in the enterprise.

Regarding services provided by the marketplace, it addresses both data consumers and data producers in a self-service manner. Data consumer services comprise things like self-service data discovery and self-service data preparation. Data producer services include, for instance, self-service data curation to define metadata on datasets as well as self-services for API-based data publishing. Marketplace services on the whole address the entire data lifecycle: data acquisition and cataloging, publishing and lineage tracking, and data preparation and exploration.

**Data applications.** Data applications refer to all kinds of applications that use data provided by the data platforms. We differentiate descriptive, diagnostic, predictive, and prescriptive data applications.<sup>15</sup> That is, data applications comprise the entire range of data analytics techniques, from reporting to machine learning. Data applications realize defined use cases, such as process performance prediction in manufacturing, for defined data consumers—for instance, a process engineer.

**Data roles.** Data roles comprise

organizational roles related to data. These roles are relevant across all layers of the data ecosystem. We focus on key roles that are of central importance for AI and data analytics in industrial enterprises—namely data owners, data stewards, data engineers, and data scientists.

Data owners<sup>1</sup> have the overall responsibility for certain kinds of data—for instance, all data on a certain product. They are assigned to the business, not IT, and are responsible for the quality, security, and compliance of this data from a business point of view. It is particularly important to define a uniform and transparent data ownership organization across the enterprise data lake and the edge data lakes and to decouple these structures from data ownership in source systems. For instance, all data on a specific product stored in the enterprise data lake should be owned by the respective business unit, to facilitate cross-process use of data.

Data stewards<sup>1</sup> manage data on behalf of data owners. They are responsible for realizing necessary policies and procedures from both business and technical points of view. To reduce the complexity and efforts of data engineering for AI, an overall data stewardship organization is needed, establishing common quality criteria and reference data models for all kinds of data. For instance, manufacturing data can be structured according to the IEC 62264 reference model<sup>20</sup> to ease data integration across different factories of the enterprise.

Data engineers and data scientists are key roles within the context of AI projects but there is no widely accepted definition—yet.<sup>28</sup> Generally, data engineers develop data pipelines to provide the data basis for further analyses by integrating and cleansing data. Building on this foundation, data scientists focus on actual data analysis by feature engineering and applying various data analytics techniques—for instance, different machine-learning algorithms—to derive insights from data.

### **From Insular AI to Industrialized AI: Addressing Challenges and Future Directions**

We are currently realizing the data ecosystem on an enterprise-scale at the manufacturer to evolve from insular AI



## **We see a major need for future research regarding functional capabilities and realization technologies for an enterprise data marketplace.**



to industrialized AI. Generally, the data ecosystem paves the way to industrialized AI by addressing the data challenges. To assess this, we analyze individual data challenges with respect to data ecosystem elements (see Table). We highlight open issues we are facing during the course of our real-world realization of the data ecosystem and point out future research directions. Further details on the realization of selected elements of the data ecosystem can be found in our most recent works.<sup>12–16</sup>

**Addressing the data management challenge.** With respect to the data management challenge, the data ecosystem is based on a comprehensive set of data platforms, namely the enterprise data lake, edge data lakes, and the enterprise data marketplace. These platforms define an enterprise data architecture for AI and data analytics, specifically addressing the aspect of data architecture. For this purpose, the enterprise data lake incorporates the enterprise data warehouse, avoiding two separate enterprise-wide data platforms and corresponding data redundancies. It is based on a unified set of data modeling guidelines and reference data models implemented by data stewards to address the aspect of data modeling. For instance, enterprise data from ERP systems is modeled using data vault modeling to enable rapid integration with sensor data from IoT devices as described in our recent work.<sup>14</sup> This enables the enterprise-wide reuse of data and data pipelines for all kinds of AI use cases across products, processes, and factories. Additionally, edge data lakes provide flexibility for use-case exploration and prototyping with only minimal guidelines, but they are restricted to local data, particularly in single factories.

The design of the platform data architecture of the enterprise data lake itself is challenging, as it must serve a huge variety of data applications, from descriptive reporting to predictive and prescriptive machine-learning applications. Particularly, defining a suitable composition of data storage and processing technologies is an open issue. According to our practical experiences, the enterprise data lake favors a polyglot approach to provide fit-for-purpose technologies for different data applications. To this end, we combine

relational database systems, NoSQL systems, and real-time event hubs following the lambda architecture paradigm as discussed in our recent work.<sup>15</sup>

Identifying suitable architecture patterns for different kinds of data applications on top of this polyglot platform constitutes a valuable future research direction for standardizing the implementation of AI use cases. In addition, organizing all data in the enterprise data lake requires an overarching structure beyond conceptual data modeling. We see data lake zones<sup>37</sup> as a promising approach necessitating substantial future research as discussed in our recent work.<sup>12</sup>

The aspect of metadata management is addressed by the data catalog as part of the enterprise data marketplace. The data catalog focuses on the acquisition, storage, and provisioning of all kinds of metadata—technical, business, and operational—across all data lakes and source systems. In this way, it enables overarching lineage analyses and data quality assessments as essential parts of AI use cases—for example, to evaluate the provenance of a dataset in the enterprise data lake. Data catalogs represent a relatively new kind of data management tool and mainly focus on the management of metadata from batch storage systems—such as relational database systems as detailed in our recent work.<sup>13</sup> Open issues particularly refer to the integrated management of metadata from batch and streaming systems, such as Apache Kafka, to realize holistic metadata management in the data ecosystem.

**Addressing the data democratiza-**

**tion challenge.** All aspects of the data democratization challenge—namely data provisioning, data engineering, and data discovery and exploration—refer to self-service and metadata management. They are addressed by the enterprise data marketplace based on the data catalog. The data catalog provides comprehensive metadata management across all data lakes and source systems of the data ecosystem. Thus, it significantly facilitates data engineering as well as data discovery and exploration for all kinds of end users by providing technical and business information on data and its sources as discussed in our recent work.<sup>16</sup> For instance, the business meaning of calculated KPIs in the enterprise data lake can be investigated, and corresponding source systems can be looked up easily in the data catalog by non-expert users.

The enterprise data marketplace also provides self-service capabilities across the entire data lifecycle for all kinds of data producers and data consumers. For instance, a process engineer in manufacturing provisions sensor data of a new machine in the enterprise data lake himself by executing a self-service workflow in the data marketplace.

Neither established tools nor sound concepts for internal enterprise data marketplaces exist, hence we are realizing the marketplace as an individual software development project. To this end, there are various realization options—for instance, using semantic technologies for modeling metadata and services.<sup>7</sup> Thus, we see a major need for future research regarding functional

capabilities and realization technologies for an enterprise data marketplace.

**Addressing the data governance challenge.** In view of the data governance challenge, the data ecosystem defines a set of key roles related to data—namely data owners, data stewards, data engineers, and data scientists. Thus, both aspects—data ownership and data stewardship—are addressed. An overarching data ownership organization across source systems and data lakes facilitates the compliant and prompt provisioning of source data for AI use cases because approvals and responsibilities for the use of data are clearly defined. Moreover, a data stewardship organization for all kinds of data significantly enhances data quality and reduces data engineering efforts by establishing reference data models and data quality criteria. At this, the data catalog supports data governance by providing KPIs for data owners and data stewards, such as the number of sources of truth for specific data sets.

A major open issue refers to the implementation of these roles within existing organizational structures. Generally, there are various data governance frameworks and maturity models in literature and practice.<sup>1,2,9,18,19,30,32,34</sup> However, they only provide high-level guidance on how to approach data governance—for example, what topics to address and what roles to define. Concrete guidelines covering how to implement data governance, considering context factors such as industry and corporate culture, are lacking—for instance, deciding when data ownership is to be organized by business unit

**Addressing data challenges by the data ecosystem and resulting future research directions.**

Data Challenges of AI	Aspects	Data Ecosystem Approach	Future Research Directions
Data Management Challenge of AI	Data Modeling	Unified data modeling concepts and reference data models in the enterprise data lake	Overall data organization in enterprise data lake—for instance, using data lake zones
	Metadata Management	Data catalog for metadata management	Integrated management of metadata from batch and streaming systems
	Data Architecture	Architecture consisting of enterprise data lake, edge data lakes, and enterprise data marketplace	Polyglot platform data architecture of enterprise data lake, including architecture patterns
Data Democratization Challenge of AI	Data Provisioning	Self-service and metadata management provided by enterprise data marketplace and data catalog	Framework of capabilities and realization technologies for an enterprise data marketplace
	Data Engineering		
	Data Discovery and Exploration		
Data Governance Challenge of AI	Data Ownership	Key roles for data owners, data stewards, data engineers, and data scientists	Implementation guidelines for data roles considering context factors—for example, corporate culture
	Data Stewardship		

or by business process.<sup>1</sup> Thus, we see a need for future research concerning context-based implementation guidelines for data roles.

**Conclusion**

Data challenges constitute the major obstacle to leveraging AI in industrial enterprises. According to our investigations of real-world industry practices, AI is currently undertaken in an insular fashion, leading to a polyglot and heterogeneous enterprise data landscape. This presents considerable challenges for systematic data management, comprehensive data democratization, and overall data governance and prevents the widespread use of AI in industrial enterprises.

To address these issues, we presented the data ecosystem for industrial enterprises as a guiding framework and overall architecture. Our assessment of the data challenges against the data ecosystem elements underlines that all data challenges are addressed—paving the way from insular AI to industrialized AI. The socio-technical character of the data ecosystem allows organizations to address both the technical aspects of the data management challenge and the organizational aspects of the data governance challenge—with defined data roles and data platforms. Furthermore, the loosely coupled and self-organizing nature of the data ecosystem with self-reliant data producers and data consumers addresses the data democratization challenge—for instance, with comprehensive self-service and metadata management provided by the enterprise data marketplace. At this, the data ecosystem is valid not only for AI but also for any kind of data analytics, as it addresses all types of data sources and all types of data applications in industrial environments. It is to be noted that the data ecosystem elements were derived from our practical findings and generalized for industrial enterprises. We encourage additional work to further refine and validate these elements.

We are currently realizing the data ecosystem at the manufacturer on an enterprise-scale and are facing various issues that indicate the need for further research. In particular, the design of an enterprise data marketplace as a novel type of data platform constitutes a valuable direction of future work.

**Acknowledgments**

The author would like to thank Jens Bockholt and Dieter Neumann for their continuous support of this work. Moreover, big thanks go to Arnold Lutsch and Eva Hoos for their valuable comments.

**References**

1. Abraham, R., Schneider, J., and Brocke, J.v. Data governance: A conceptual framework, structured review, and research agenda. *Intern. J. of Information Management* 49 (2019), 424–438.
2. Ballard, C., Compert, C., Jesionowski, T., Milman, I., Plants, B., Rosen, B., and Smith, H. Information governance principles and practices for a big data landscape. *IBM* (2014).
3. Ballou, D.P. and Tayi, G.K. Enhancing data quality in data warehouse environments. *Communications of the ACM* 42, 1 (1999), 73–78.
4. Cao, L. Data science: A comprehensive overview. *ACM Computing Surveys* 50, 3 (2017), 1–42.
5. Chu, X., Ilyas, I.F., Krishnan, S., and Wang, J. Data cleaning: Overview and emerging challenges. In *Proceedings of the Intern. Conf. on Management of Data (SIGMOD)*, ACM, New York (2016), 2201–2206.
6. Cui, Y., Kara, S., and Chan, K.C. Manufacturing big data ecosystem: A systematic literature review. *Robotics and Computer Integrated Manufacturing* 62 (2020) Article 101861.
7. Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., and Bartolucci, A. The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics* 108, 1 (2016), 441–455.
8. Davenport, T.H. and Ronanki, R. Artificial intelligence for the real world. *Harvard Business Review* 96, 1 (2018), 108–116.
9. The DGI data governance framework. The Data Governance Institute (2020).
10. Everitt, T. and Hutter, M. Universal artificial intelligence. Practical agents and fundamental challenges. *Foundations of Trusted Autonomy*. H. Abbass, J. Scholz, and D. Reid, eds. Springer. (2018) 15–46.
11. Gessert, F., Wingerath, W., Friedrich, S., and Ritter, N. NoSQL database systems: A survey and decision guidance. *Computer Science—Research and Development* 32, 3–4 (2016), 353–365.
12. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. A zone reference model for enterprise-grade data lake management. In *Proceedings of the IEEE Enterprise Distributed Object Computing Conf. (EDOC)*, IEEE, Piscataway, New Jersey (2020), 57–66.
13. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. Leveraging the data lake: Current state and challenges. In *Proceedings of the Intern. Conf. on Big Data Analytics and Knowledge Discovery (DaWaK)*, Springer, Cham, (2019), 179–188.
14. Giebler, C., Gröger, C., Hoos, E., Schwarz, H., and Mitschang, B. Modeling data lakes with data vault: Practical experiences, assessment, and lessons learned. In *Proceedings of the Intern. Conf. on Conceptual Modeling (ER)*, Springer, Cham, (2019), 63–77.
15. Gröger, C. Building an Industry 4.0 analytics platform. *Datenbank-Spektrum* 18, 1 (2018), 5–14.
16. Gröger, C. and Hoos, E. Ganzheitliches metadatenmanagement im data lake: Anforderungen, IT-werkzeuge und herausforderungen in der praxis. In *Proceedings Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW)*, Gesellschaft für Informatik, Bonn, (2019), 435–452.
17. Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Amsterdam, (2012).
18. Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., and Smith, E. *DAMA-DMBOK: Data Management Body of Knowledge*. Technics Publications, New Jersey, (2017).
19. Holistic data governance: A framework for competitive advantage. Informatica (2017).
20. IEC 62264-2:2015. Enterprise-control system integration—Part 2: Objects and attributes for enterprise-control system integration. International Organization for Standardization (2015).

21. Ismail, A., Truong, H.-L., and Kastner, W. Manufacturing process data analysis pipelines: A requirements analysis and survey. *Journal of Big Data* 6, 1 (2019), 1–26.
22. Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., and Eschert, T. Industrial Internet of Things and cyber manufacturing systems. *Industrial Internet of Things*. S. Jeschke, C. Brecher, H. Song, and D. Rawat, eds. Springer, (2017), 3–19.
23. Kimball, R. and Ross, M. *The Data Warehouse Toolkit. The Definitive Guide to Dimensional Modeling*. Wiley, Indianapolis, (2013).
24. Laudon, K.C. and Laudon, J.P. *Management Information Systems. Managing the Digital Firm*. Pearson Education, Harlow, (2018).
25. Lee, J., Davari, H., Singh, J., and Pandhare, V. Industrial artificial intelligence for Industry 4.0-based manufacturing systems. *Manufacturing Letters* 18, (2018), 20–23.
26. Linstedt, D. and Olschmke, M. *Building a Scalable Data Warehouse with Data Vault 2.0*. Morgan Kaufmann, Waltham, (2016).
27. Loucks, J., Davenport, T.H., and Schatsky, D. State of AI in the enterprise, 2<sup>nd</sup> edition. Deloitte, (2018).
28. Lyon, L. and Mattern, E. Education for real-world data science roles (part 2): A translational approach to curriculum development. *Intern. J. of Digital Curation* 11, 2 (2016), 13–26.
29. Mathis, C. Data lakes. *Datenbank-Spektrum* 17, 3 (2017), 289–293.
30. Morabito, V. *Big Data and Analytics*. Springer, Cham, (2015).
31. Oliveira, M.I.S., Fatima Barros Lima, G.d., and Loscio, B.F. Investigations into data ecosystems: A systematic mapping study. *Knowledge and Information Systems* 61, 2 (2019), 589–630.
32. Plotkin, D. *Data Stewardship*. Morgan Kaufmann, (2014).
33. Reggi, L. and Dawes, S. Open government data ecosystems: Linking transparency for innovation with transparency for participation and accountability. In *Proceedings of the Intern. Conf. on Electronic Government (EGOV)*, Springer, Cham, (2016), 74–86.
34. The SAS data governance framework: A blueprint for success. SAS (2018).
35. Schaeffer, E., Wahrenndorff, M., Narsalay, R.M., Gupta, A., and Hobräck, O. Turning possibility into productivity. *Accenture* (2018).
36. Schomm, F., Stahl, F., and Vossen, G. Marketplaces for data: An initial survey. *ACM SIGMOD Record* 42, 1 (2013), 15–26.
37. Sharma, B. *Architecting Data Lakes*. O'Reilly, Sebastopol, CA, (2018).
38. Smith, G., Ofc, H.A., and Sandberg, J. Digital service innovation from open data: Exploring the value proposition of an open data marketplace. In *Proceedings of the Hawaii Intern. Conf. on System Sciences (HICSS)*, IEEE, Piscataway, New Jersey, (2016), 1277–1286.
39. Taleb, I., Serhani, M.A., and Dssouli, R. Big data quality: A survey. In *Proceedings of the IEEE Intern. Congress on Big Data, IEEE*, Piscataway, New Jersey, (2018), 166–173.
40. Yang, Y., Pan, L., Ma, J., Yang, R., Zhu, Y., Yang, Y., and Zang, L. A high-performance deep-learning algorithm for the automated optical inspection of laser welding. *J. of Applied Sciences* 10, 3 (2020), 1–11.
41. Zeng, J. and Glaister, K.W. Value creation from big data: Looking inside the black box. *Strategic Organization* 16, 2 (2018), 105–140.

**Christoph Gröger** (christoph.groeger@de.bosch.com) is enterprise architect for data analytics at Bosch and a senior technical professional in Bosch's global data strategy team in Stuttgart, Germany.

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



Watch the author discuss this work in the exclusive *Communications* video.  
<https://caem.acm.org/videos/no-ai-without-data>

## Automatic map inference, data refinement, and machine-assisted map editing promises more accurate map datasets.

BY FAVYEN BASTANI, SONGTAO HE, SATVAT JAGWANI, EDWARD PARK, SOFIANE ABBAR, MOHAMMAD ALIZADEH, HARI BALAKRISHNAN, SANJAY CHAWLA, SAM MADDEN, AND MOHAMMAD AMIN SADEGHI

# Inferring and Improving Street Maps with Data-Driven Automation

STREET MAPS HELP to inform a wide range of decisions. Drivers, cyclists, and pedestrians use them for search and navigation. Rescue workers responding to disasters such as hurricanes, tsunamis, and earthquakes rely on street maps to understand where people are and to locate individual buildings.<sup>23</sup>

Transportation researchers consult street maps to conduct transportation studies, such as analyzing pedestrian accessibility to public transport.<sup>25</sup> Indeed, with the need for accurate street maps growing in importance, companies are spending hundreds of millions of dollars to map roads globally.<sup>a</sup>

However, street maps are incomplete or lag behind new construction in many parts of the world. In rural Indonesia, for example, entire groups of

<sup>a</sup> For example: Korosec, K. Uber will spend \$500 million on mapping to diverge from Google. *Fortune* (May 7, 2016).

### » key insights

- Digital street maps cost hundreds of millions of dollars annually to create and maintain, but updates lag months behind new road and building construction.
- Satellite imagery and GPS data captured by road users are promising data sources for automating the map update process. Machine learning can infer road and building positions from these sources.
- To update maps with minimal error and cost, we built Mapster, which features a machine-assisted map editing framework and end-to-end machine-learning methods that directly infer vector road networks from raw data sources.

villages are missing from OpenStreetMap, a popular open map dataset.<sup>3</sup> In many of these villages, the closest mapped road is miles away. In Qatar, construction of new infrastructure has boomed in preparation for the FIFA World Cup 2022. But the rapid pace of construction means that it often takes a year for digital maps to reflect new roads and buildings.<sup>b</sup> Even in countries such as the U.S., where significant investment has been made in digital maps, construction and road closures often take days or weeks to be incorporated into map datasets.

These problems arise because today's processes for creating and maintaining street maps are extremely labor-intensive. Modern street map editing tools allow users to trace and annotate roads and buildings on a map canvas overlaid on relevant data sources, so that users can effectively edit the map while consulting the data. These data sources include satellite imagery, aerial imagery, and GPS trajectories (which consist of sequences of GPS positions captured from moving vehicles). Although the data presented by these tools helps users to update a map dataset, the manual tracing and annotation process is cumbersome and a major bottleneck in map maintenance.

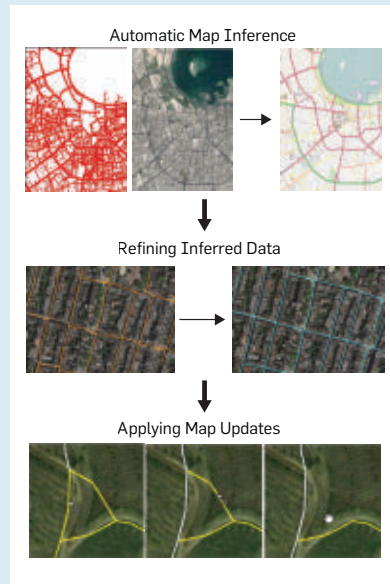
Over the past decade, many *automatic map inference* systems have been proposed to automatically extract information from these data sources at scale. Several approaches develop unsupervised clustering and density-thresholding algorithms to construct road networks from GPS trajectory datasets.<sup>1,4,7,16,21</sup> Others apply machine-learning methods to process satellite imagery and extract road networks,<sup>19,22</sup> building polygons,<sup>13,24</sup> and road attribute annotations—for example, the number of lanes, presence of cycling lanes, or presence of street parking on a road.<sup>6,18</sup>

However, automatic map inference has failed to gain traction in practice due to two key limitations. First, existing inference methods have high error rates (low precision), which manifest in noisy outputs containing incorrect

<sup>b</sup> An example of a subdivision in Doha, Qatar that was missing from maps for years is detailed at <https://productforums.google.com/forum/#!topic/maps/dwtCso9owlU>.

**Figure 1. An overview of the Mapster map editing system.**

First, we infer the road network from satellite imagery and GPS data. Then, we transform the map to make it look more realistic. Finally, we have an interactive system to apply map updates.



roads, buildings, and attribute annotations. Second, although prior work has shown how to detect roads and buildings from data sources, the challenge of leveraging this information to update real street map datasets has not been addressed. We argue that even with lower error rates, the quality of outputs from automatic approaches is below that of manually curated street map datasets, and semi-automation is needed to efficiently leverage automatic map inference to accelerate the map maintenance process.

At Massachusetts Institute of Technology (MIT) and Qatar Computer Research Institute (QCRI), we have developed several algorithms and approaches to address these challenges,<sup>2,3,14</sup> which we combined into a new system called *Mapster*. Mapster is a human-in-the-loop street map editing system that incorporates three components to accelerate the mapping process over traditional tools and workflows: high-precision automatic map inference, data refinement, and machine-assisted map editing.

First, Mapster applies automatic map inference algorithms to extract initial estimates from raw data sources. Although these estimates are noisy,

we minimize errors by applying two novel approaches that replace heuristic, error-prone post-processing steps present in prior work with end-to-end machine learning and other more robust methods: *iterative tracing for road network inference* and *graph network annotation for attribute inference*.

Second, Mapster refines the noisy estimates from map inference into map update proposals by removing several types of errors and reducing noise. To do so, we apply conditional generative adversarial networks (cGANs) trained to transform the noisy estimates into refined outputs that are more consistent with human-annotated data.

Finally, a machine-assisted map editing framework enables the rapid, semi-automated incorporation of these proposed updates into the street map dataset. This editing tool addresses the problem of leveraging inferred roads, buildings, and attribute annotations to update existing street map datasets.

Figure 1 summarizes the interactions between these components. We included links to two videos demonstrating the execution of Mapster, along with a link to Mapster's source code (which we have released as free software).<sup>c</sup>

In this article, we first introduce our automatic map inference approaches for inferring roads and road attributes, which achieve substantially higher precision than prior work. Then, we detail our data refinement strategy, which applies adversarial learning to improve the quality of inferred road networks. Finally, we discuss our machine-assisted map editor, which incorporates novel techniques to maximize user productivity in updating street maps. We conclude with a discussion of future work.

### Automatically Inferring Road Networks

Given a base road network, which may be empty or may correspond to the roads in the current street map dataset, the goal of road network inference is to leverage GPS trajectory data and satellite imagery to produce a road network

<sup>c</sup> Iterative tracing in action: [https://youtu.be/3\\_AE2Qn-Rdg](https://youtu.be/3_AE2Qn-Rdg). Machine-assisted map editing: <https://youtu.be/i-6nbuuX6NY>. Mapster source code: <https://github.com/mitroad-maps>.



map that covers roads not contained in the base map. The road network map is represented as a graph, where vertices are annotated with spatial longitude-latitude coordinates and edges correspond to straight-line road segments.

Broadly, prior approaches infer roads by dividing the space into a 2D grid, classifying whether each grid cell contains a road, and connecting cells together to form edges. Figure 2(a) summarizes this strategy. For satellite imagery, recent work obtains the per-cell classification by applying deep convolutional neural networks (CNNs) that segment the imagery, transforming the input imagery into a single-channel image that indicates the neural network’s confidence that there is a road at each pixel.<sup>9,10,17</sup> For GPS trajectories, several approaches perform the classification based on the number of GPS trajectories passing through each cell.<sup>5,8,11,20</sup>

However, we find that these methods exhibit low accuracy in practice when faced with challenges such as noisy data and complex road topology. Figure 2(b) shows the output of prior work around a major highway junction in Chicago. Noise in the per-cell classification estimates is amplified when we connect cells together to draw edges, resulting in garbled road network maps.

Indeed, both GPS trajectory data and satellite imagery exhibit several types of noise that make robust identification of roads challenging. While GPS samples in open areas are typically accurate to four meters, in practice, due to high-rise buildings and reflection, GPS readings may be as far off as tens of meters. Correcting this error is difficult because errors are often spatially correlated—multiple GPS readings at the same location may be offset in the same way as they encounter the

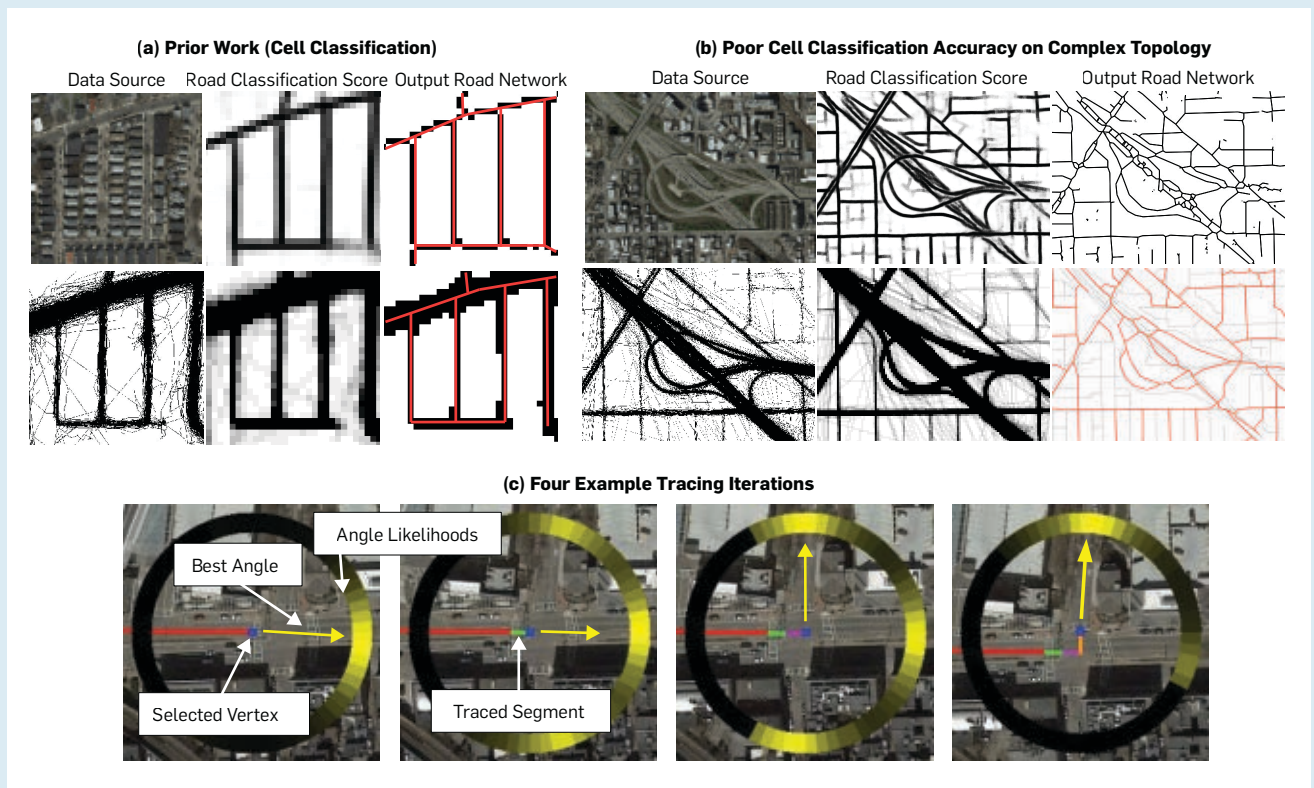
same reflection and distortion issues. Similarly, roads in satellite imagery are frequently occluded by trees, buildings, and shadows. Furthermore, distinguishing roads and buildings from non-road trails and surface structures in imagery is often nontrivial.

To substantially improve precision, we adopt an iterative road-tracing approach in lieu of the per-cell classification strategy. Our iterative tracing method mimics the gradual tracing process that human map editors use to create road network maps, thereby eliminating the need for the heuristic post-processing steps that prior work applies to draw edges based on cell classification outputs.

Iterative tracing begins with the base map, and on each iteration, it adds a single road segment (one edge) to the map. To decide where to position this segment, it uses the data

**Figure 2. Old and new mapping approaches.**

(a) Prior automatic mapping approaches operating on satellite imagery (top) and GPS trajectories (bottom). In the center, we show the per-cell classification scores, and on the right, the result of thresholding and cell connection. (b) These approaches produce noisy outputs around complex road topology such as highway interchanges. (c) Iterative tracing on satellite imagery. A CNN predicts the likelihood that there is a road at each of 64 angles from a vertex; higher likelihoods from the blue vertex are shown in yellow in the outer circle, and lower likelihoods are shown in black. Iterative tracing gradually expands the coverage of the map: on each iteration, it adds a segment corresponding to the highest likelihood vertex and direction.



source to compute two values for each vertex in the portion of the map traced so far: (a) a confidence that an unmapped road intersects the vertex, and (b) the most likely angular direction of the unmapped road. Then, it selects the vertex with the highest confidence and adds a segment in the corresponding direction. This procedure is repeated, adding one segment

at a time to the map, until the highest confidence for the presence of an unmapped road falls below a threshold. We illustrate the iterative tracing procedure in Figure 2(c).

We develop different approaches to compute the unmapped road confidence and direction from satellite imagery<sup>3</sup> and from GPS trajectories.<sup>14</sup> With satellite imagery, we compute

these through a deep neural network. We develop a CNN model that inputs a window of satellite imagery around a vertex of the road network, along with an additional channel containing a representation of the road network that has been traced so far. We train the CNN to output the likelihood that an unmapped road intersects the vertex, and the most likely angular direction of this unmapped road.

With GPS trajectories, we compute the values at a vertex based on the peak direction of trajectories that pass through the vertex. We identify all trajectories that pass through the vertex and construct a smoothed polar histogram over the directions followed by those trajectories after they move away from the vertex. Then, we apply a peak-finding algorithm to identify peaks in the histogram that have not been explored earlier in the tracing process. We select the peak direction and measure confidence in terms of the volume of trajectories that match the peak direction.

Often, both satellite imagery and GPS trajectory data may be available in a region requiring road network inference. To reduce errors and improve the map quality, we develop a two-stage approach that leverages both data sources when inferring road networks. In the first stage, we run iterative tracing using GPS data to infer segments along high-throughput roadways. Because these roads have high traffic volume, they are covered by large numbers of GPS trajectories, so they can be accurately traced with GPS data. At the same time, junctions along high-throughput roads (especially controlled-access highways) are generally more complex, often involving roundabouts and overpasses. These features make tracing based on satellite imagery challenging.

In the second stage, we fill in gaps in the road network with lower-throughput residential and service roads that were missed in the first stage by tracing with satellite imagery. These roads have simple topologies and are covered by few GPS trajectories, making imagery a preferred data source.

We evaluate our method by comparing road networks inferred through iterative tracing against those inferred by prior cell-classification approaches. Figure 3 shows qualitative results in Boston, Chicago, Salt Lake City, and

**Figure 3. Qualitative results comparing iterative tracing to prior cell classification approaches shows (a) road networks inferred from satellite imagery, and (b) road networks inferred from GPS trajectories. We show inferred roads in the foreground and OpenStreetMap data in the background.**



Los Angeles. We use 60-cm/pixel satellite imagery from the United States Geological Survey (USGS) and Google Earth, which is available in urban areas across the world, and GPS trajectory data captured at 1 Hz from private cars. In contrast to cell classification, iterative tracing from satellite imagery robustly infers roads despite occlusion by buildings and shadows in dense urban areas. In lower-density areas such as Salt Lake City, iterative tracing performs comparably to prior work.

Cell classification from GPS trajectories produces noisy outputs at crucial but complex map features, such as highway interchanges. Despite the intricate connections at these features, iterative tracing accurately maps the interchanges.

We show quantitative results.<sup>3,14</sup> The execution time of iterative tracing is practical. With satellite imagery, the asymptotic complexity is  $O(l)$ , where  $l$  is the total length of the inferred roads. For example, inferring roads in a 15-km<sup>2</sup> urban area requires 16 minutes on an NVIDIA Tesla V100 GPU with iterative tracing; although slower than cell classification (which completes in two minutes), it is nevertheless efficient even for large road networks, and tracing can be parallelized across multiple machines. With GPS trajectories, the asymptotic complexity is  $O(l \cdot d)$ , where  $l$  is the total length of the inferred roads and  $d$  is the average number of GPS points explored at each iteration. For example, iterative tracing takes 110 minutes to process 2.5 million GPS points in a 16-km<sup>2</sup> area on an AWS C5.9xlarge instance with 15% CPU utilization.

### Inferring Road Attributes

Modern navigation systems use more than just the road topology—they also use a number of road metadata attributes, such as the number of lanes, presence of cycling lanes, or the presence of street parking along a road. As a result, inferring these attributes is an important part of the Mapster system.

Prior work in road-attribute inference applies an image-classification approach that trains a CNN to predict the road attributes given a window of satellite imagery around some location along the road. Then, the local prediction at each location is post-processed through



**Mapster is a street map editing system that accelerates the mapping process over traditional tools and workflows with high-precision automatic map inference, data refinement, and machine-assisted map editing.**



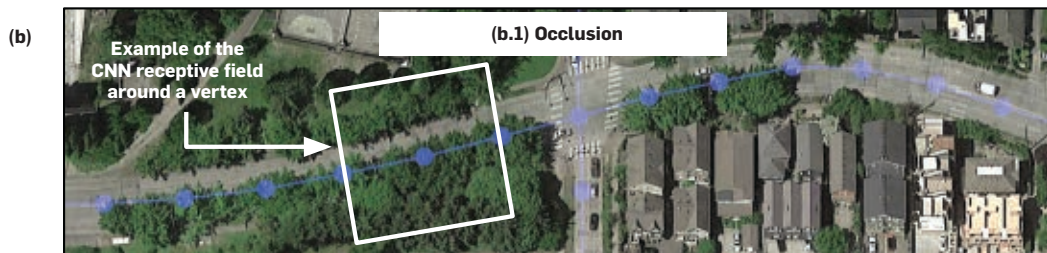
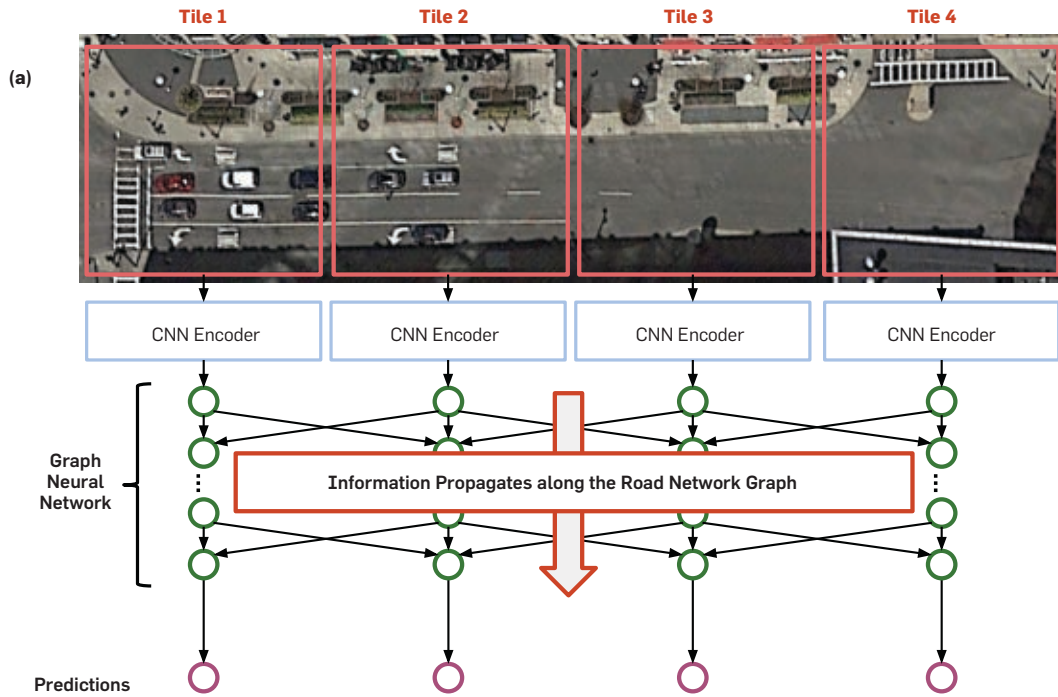
a global inference phase to remove scattered errors—for example, inference in a Markov Random Field (MRF).

This global inference phase is necessary because the CNN prediction at each location is often erroneous due to the *limited receptive field* of the CNN—in many cases, local information in the input window of satellite imagery is not sufficient to make a correct prediction. For example, in Figure 4(b.1), the road on the left side is occluded by trees. If the CNN inputs a window only from the left part of the road, it will be unable to correctly determine the number of lanes. The global inference phase corrects these errors in the CNN outputs by accounting for all the predictions along the road, as well as prior knowledge—for example, road attributes such as the number of lanes should generally be homogeneous along the road instead of varying frequently.

However, we find that global inference postprocessing is often error prone. For example, consider Figure 4(b.2), where the lane count changes from four to five near an intersection. The image classifier outputs partially incorrect labels. The global inference phase fails to correct this error; because it only accounts for predictions from the image classifier, it cannot determine whether the number of lanes indeed changes or is simply an error of the image classifier. This limitation is caused by the *information barrier* induced by the separation of local classification and global inference; the global inference phase can only use the image classifier's prediction as input, but not other important information, such as whether trees occlude the road or whether the road width changes.

To break the information barrier, we develop a hybrid neural network architecture, as in Figure 4(a), which combines a CNN with a graph neural network (GNN). The CNN extracts local features from each segment along a road. Then, the GNN propagates these features along the road network. End-to-end training of the combined CNN and GNN model is key to the success of the method: rather than rely on hand-crafted, error-prone, post-processing heuristics that only operate over CNN predictions, our method instead *learns*

**Figure 4. (a) The hybrid neural-network architecture proposed in this work. (b) Examples on inferring the number of lanes. In each image, blue lines show the road network. The number of lanes predicted by the CNN Image Classifier and Mapster on each segment are shown along the bottom of each figure. Output numbers in green are correct predictions while red numbers are incorrect predictions.**



CNN Image Classifier Only:

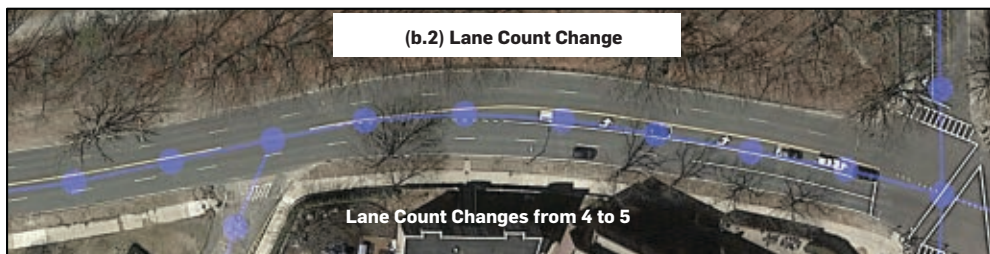
4	1	1	1	2	1	-	4	4	2	4	4	4
---	---	---	---	---	---	---	---	---	---	---	---	---

CNN Image Classifier + Post-Processing (MRF):

4	4	4	4	4	4	-	4	4	4	4	4	4
---	---	---	---	---	---	---	---	---	---	---	---	---

Mapster's CNN+GNN solution":

4	4	4	4	4	4	-	4	4	4	4	4	4
---	---	---	---	---	---	---	---	---	---	---	---	---



CNN Image Classifier Only:

4	4	-	4	4	5	5	5	4	-
---	---	---	---	---	---	---	---	---	---

CNN Image Classifier + Post-Processing (MRF):

4	4	-	4	4	4	4	4	4	-
---	---	---	---	---	---	---	---	---	---

Mapster's CNN+GNN solution":

4	4	-	4	4	5	5	5	5	-
---	---	---	---	---	---	---	---	---	---

the post-processing rules as well as the required CNN features directly from the data. The use of the GNN in our system eliminates the *receptive field limitation* of local CNN image classifiers, and the combination of CNN and GNN eliminates the *information barrier* present in prior work. This allows the model to learn complex inductive rules that make it more robust in the face of challenges, such as occlusion in satellite imagery and the partial disappearance of important information, as seen in Figure 4(b).

### Refining Inferred Data

Although iterative tracing improves substantially over prior grid-cell classification approaches, it nevertheless creates road networks with noisy features that clearly distinguish them from human-drawn maps. Iterative tracing is particularly effective at capturing road network topology but leaves geometrical abnormalities, such as the examples of off-center junction vertices and noisy road curvature in Figure 5(a).

To refine the map and rectify these abnormalities, we use a model—based on cGANs<sup>12</sup>—to learn the realistic road appearance. These networks learn to realistically reproduce complex, image-to-image transformations and have been successfully applied to many tasks, including adding color to black-and-white images and transforming photos taken in daylight to plausible nighttime photos of the same scene.<sup>15</sup>

An obvious transformation to learn for map inference is transforming satellite imagery or representations of GPS trajectories into road networks, where the output image contains lines corresponding to road segments. However, we find that the learning problem is too difficult under this strategy, and the cGAN model fails to learn to robustly identify roads. Instead, it primarily learns to produce arbitrary lines that resemble lines in the ground-truth road-network representation.

Thus, our cGAN model instead inputs not only satellite imagery or GPS trajectory data, but also a representation of the road network produced by iterative tracing. It outputs a refined road network representation where abnormalities in the input network

have been corrected. By providing this initial road network, we reduce the complexity of the transformation and thereby assist the cGAN to learn the transformation, especially early in the training process. Incorporating the initial road network representation derived from iterative tracing into the cGAN was a crucial insight that made training the adversarial model feasible.

Our cGAN architecture consists of two components: a *generator* that produces refined road networks given an initial road network and a data source, and a *discriminator* that learns to dis-

tinguish refinements made by the generator from road networks in the ground truth dataset. This network is adversarial because we train the generator and discriminator with opposing loss functions: the discriminator minimizes its classification error at distinguishing ground truth and generated (refined) road networks. In contrast, the generator learns by maximizing the discriminator's classification error. Thus, in effect, we train the generator by having it learn to fool the discriminator into classifying its generated road network as ground truth.

**Figure 5. (a) Road networks before refinement, in orange, contain geometric abnormalities. The refined road networks, shown in blue, clean up these noisy features. (b) Our human-in-the-loop map-editing system employs a pruning algorithm to eliminate residential and service roads and focus the user on adding major roads to the map. Pruned roads are shown in purple and the remaining roads in yellow.**



The initial road network provided to the generator enables the cGAN model to learn to produce realistic road networks. At first, the generator simply copies the road network from its input to its output to deceive the discriminator. However, once the discriminator learns to better distinguish the iterative-tracing road networks from hand-drawn, ground truth roads, the generator begins making small adjustments to the roads so that they appear to be hand-drawn. As training continues, these adjustments become more robust and complete.


Figure 5(a) shows the outputs of our cGAN in blue, both on the geometrical abnormalities introduced earlier and on a larger region in Minneapolis, MN. Again, we use 60-cm/pixel satellite imagery for this evaluation. While refinement does not substantially alter the topology of the road network, the cGAN improves the geometry so that inferred roads resemble hand-mapped roads. These geometry improvements help to reduce the work needed to integrate inferred data into the street map.

### Machine-Assisted Map Editing


To improve street map datasets, the inferred road network derived from iterative tracing and refinement steps must be incorporated into the existing road network map. Fully automated integration of the inferred road network is impractical: inferred roads may include errors even after refinement, so adding all the inferred roads to the map dataset would degrade the its precision.

Instead, we developed a human-in-the-loop map-editing framework that enables human map editors to quickly validate automatically inferred data.<sup>2</sup> On initialization, our machine-assisted map editor builds an overlay containing the inferred road segments. Users interact with the overlay by left- and right-clicking to mark segments as correct or incorrect. Thus, rather than trace roads through a series of repeated clicks along the road, when a correctly inferred segment already covers the road, users of the machine-assisted editor can rapidly add the road to the map with a single click on the inferred segment in the overlay.

Our map editor has two additional



**Modern navigation systems use more than just the road topology—they use road metadata attributes, such as the number of lanes, presence of cycling lanes, or the presence of street parking along a road. Inferring these attributes is an important part of Mapster.**



features to improve validation speed. First, the interface includes a “prune” button that executes a shortest-path-based pruning algorithm to eliminate minor residential and service roads from the overlay, leaving only major arterial roads. This functionality is especially useful when mapping areas where the existing road network in the street map dataset has low coverage. In these areas, adding every missing road to the map may require substantial effort, but the quality of the map could be improved significantly just by incorporating major roads. The pruning algorithm is effective at helping users focus on mapping these unmapped major roads by reducing information overload. We show an example pruning result in Figure 5(b). Purple segments are pruned, leaving the yellow segments that correspond to major inferred roads.

Pruning is most useful in low-coverage areas, but for high-coverage areas, we developed a teleport button that pans users to a connected component of inferred roads. In high-coverage areas, only a small number of roads are missing from the map, and identifying an unmapped road requires users to painstakingly scan the imagery one tile at a time. The teleport button eliminates this need, allowing users to jump to a group of missing roads and immediately begin mapping them.

### Future Work

Although Mapster greatly reduces map creation workload and maintenance, automatically inferred street maps still have more errors than maps created by professional mapmakers. Narrowing this gap requires advances in machine-learning approaches for automatic map inference. Next, we detail several promising avenues for improving inference performance.

First, better neural-network architectures can improve the performance of automatic mapping. So far, the design of the neural-network architectures in Mapster are mostly inspired by best practices in general computer vision tasks, such as image segmentation and image classification. However, map inference tasks have unique characteristics, such as the strong spatial correlation in satellite imagery. Thus, improved neural-network architectures that are specialized

for mapmaking tasks may yield higher accuracy. For example, instead of using raw images or GPS traces as input, node and edge embeddings garnered from unsupervised tasks on extremely large map datasets could be used.


Second, end-to-end loss functions are a promising avenue to directly learn desired properties of the output road network. However, these desired properties, such as high precision and high recall over roads, generally can only be computed from the road network graph. As a result, the objectives are nondifferentiable since the graph can only be extracted from the probabilistic output of a machine-learning model through non-differentiable functions. Thus, both prior work and our iterative-tracing approach learn to build a graph indirectly: prior work minimizes the per-cell classification error, and in iterative tracing on satellite imagery, we minimize the difference between the predicted angle of an untraced road and the correct angle on each tracing step. Nevertheless, end-to-end training has been shown to improve performance in other machine-learning tasks and could improve the accuracy of automatic map creation and maintenance.

Several techniques have been proposed for optimizing non-differentiable metrics. For example, reinforcement-learning techniques can search for optimal policies under non-differentiable rewards. Alternatively, it may be possible to train an additional neural network, which takes the intermediate map representation—for example, road segmentation—as input and predicts the value of evaluation metrics. This additional neural network acts as a differentiable approximation of the evaluation metrics and can be leveraged to train a model to infer road networks that score highly on the metric.

In addition to potential improvements in machine-learning techniques, incorporating new data sources would also extend the capability of Mapster. Two particularly promising data sources are drone imagery and dashboard-camera video. Drone imagery enables on-demand image sensing—for instance, if we find the road structure in a region is unclear from satellite imagery and GPS data, we can reactively assign drones to collect aerial images over that region. Video from

dashboard cameras enables inferring several additional street map features, such as street names, business names, road signs, and road markers.

## Conclusion

The world has approximately 64 million kilometers of roads and the road network is growing at a rapid pace as major countries such as China, India, Indonesia, and Qatar gather economic momentum. Street maps are important. However, creating and maintaining street maps is very expensive and involves labor-intensive processes. As a result, although a lot of effort and money has been spent in maintaining street maps, today's street maps are still imperfect and are frequently either incomplete or lag behind new construction. Mapster is a holistic approach for applying automation to reduce the work needed to maintain street maps. By incorporating automatic map inference with data refinement and a machine-assisted map editor, Mapster makes automation practical for map editing, and enables the curation of map datasets that are more complete and up to date at less cost. 

## References

- Ahmed, M., Karagiorgou, S., Pfoser, D., and Wenk, C. A comparison and evaluation of map construction algorithms using vehicle tracking data. *GeoInformatica* 19, 3 (2015), 601–632.
- Bastani, F., He, S., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., and Madden, S. Machine-assisted map editing. In *Proceedings of the 26th ACM SIGSPATIAL Intern. Conf. on Advances in Geographic Information Systems* (2018), 23–32.
- Bastani, F., He, S., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., Madden, S., and DeWitt, D. RoadTracer: Automatic extraction of road networks from aerial images. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition* (2018), 4720–4728.
- Biagioni, J. and Eriksson, J. Inferring road maps from global positioning system traces: Survey and comparative evaluation. *J. of the Transportation Research Board* 2291, 1 (2012).
- Biagioni, J. and Eriksson, J. Map inference in the face of noise and disparity. In *Proceedings of the 20th Intern. Conf. on Advances in Geographic Information Systems*, ACM, (2012), 79–88.
- Cadamuro, G., Muhebwa, A., and Taneja, J. Assigning a grade: Accurate measurement of road quality using satellite measurements. In *Proceedings of NIPS Workshop on Machine Learning for the Developing World* (2018).
- Cao, L. and Krumm, J. From GPS traces to a routable road map. In *ACM SIGSPATIAL* (2009), 3–12.
- Chen, C. and Cheng, Y. Roads digital map generation with multi-track GPS data. In *Proceedings of the 2008 Intern. Workshop on Geoscience and Remote Sensing*.
- Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., and Pan, C. Automatic road detection and centerline extraction via cascaded end-to-end convolutional neural network. *IEEE Trans. on Geoscience and Remote Sensing* 55, 6 (2017), 3322–3337.
- Costea, D. and Leordeanu, M. Aerial image geolocalization from recognition and matching of roads and intersections. In *Proceedings of the British Machine Vision Conf.* (2016).
- Davies, J.J., Beresford, A.R., and Hopper, A. Scalable, distributed, real-time map generation. *IEEE Pervasive Computing* 5, 4 (2006), 47–54.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems* (2014), 2672–2680.
- Hamaguchi, R. and Hikosaka, S. Building detection from satellite imagery using ensemble of size-specific detectors. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2018).
- He, S., Bastani, F., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., and Madden, S. RoadRunner: Improving the precision of road network inference from GPS trajectories. In *Proceedings of the 26th ACM SIGSPATIAL Intern. Conf. on Advances in Geographic Information Systems* (2018).
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A.A. Image-to-image translation with conditional adversarial nets (2016).
- Liu, X., Biagioni, J., Eriksson, J., Wang, Y., Forman, G., and Zhu, Y. Mining large-scale, sparse GPS traces for map inference: Comparison of approaches. In *Proceedings of the 18th ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data Mining* (2012).
- Mattyus, G., Luo, W., and Urtasun, R. DeepRoadMapper: Extracting road topology from aerial images. In *Proceedings of the IEEE Intern. Conf. on Computer Vision* (2017), 3438–3446.
- Mattyus, G., Wang, S., Fidler, S., and Urtasun, R. Enhancing road maps by parsing aerial images around the world. In *Proceedings of the IEEE Intern. Conf. on Computer Vision* (2015), 1689–1697.
- Mnih, V. and Hinton, G.E. Learning to detect roads in high-resolution aerial images. In *European Conf. on Computer Vision*, Springer (2010), 210–223.
- Shi, W., Shen, S., and Lu, Y. Automatic generation of road network map from massive GPS, vehicle trajectories. In *Proceedings of the 12th Intern. IEEE Conf. on Intelligent Transportation Systems* (2009).
- Stanojevic, R., Abbar, S., Thirumuruganathan, S., Chawla, S., Filali, F., and Aleimat, A. Robust road map inference through network alignment of trajectories. In *Proceedings of the 2018 SIAM Intern. Conf. on Data Mining* (2018).
- Wegner, J.D., Montoya-Zegarra, J.A., and Schindler, K. Road networks as collections of minimum cost paths. *ISPRS J. of Photogrammetry and Remote Sensing* 108 (2015), 128–137.
- Yin, A. A mapathon to pinpoint areas hardest hit in Puerto Rico. *The New York Times* (October 2017).
- Zhao, K., Kang, J., Jung, J., and Sohn, G. Building extraction from satellite images using mask r-cnn with building boundary regularization. In *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops* (June 2018).
- Zielstra, D. and Hochmair, H.H. Comparative study of pedestrian accessibility to transit stations using free and proprietary network data. *J. of the Transportation Research Board* 2217, 1 (2011), 145–152.

**Favyen Bastani** is a Ph.D. student at MIT Computer Science and Artificial Intelligence Lab (CSAIL), CSAIL, Cambridge, MA, USA.

**Songtao He** is a Ph.D. student at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Satvat Jagwani** is a master's student at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Edward Park** is a master's student at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Sofiane Abbar** is a machine-learning software engineer at Facebook, U.K.

**Mohammad Alizadeh** is an associate professor of Electrical Engineering and Computer Science at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Hari Balakrishnan** is a professor of Electrical Engineering and Computer Science at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Sanjay Chawla** is the research director of the Data Analytics department at Qatar Computer Research Institute (QCRI), Doha, Qatar.

**Sam Madden** is a professor of Electrical Engineering and Computer Science at MIT CSAIL, CSAIL, Cambridge, MA, USA.

**Mohammad Amin Sadeghi** is a research scientist at QCRI, Doha, Qatar.



## ACM BOOKS

### Collection II

*Sir Tony Hoare has had an enormous influence on computer science, from the Quicksort algorithm to the science of software development, concurrency, and program verification. His contributions have been widely recognized: He was awarded the Turing Award in 1980, the Kyoto Prize from the Inamori Foundation in 2000, and was knighted for "services to education and computer science" by Queen Elizabeth II of England in 2000.*

*This book presents the essence of his various works—the quest for effective abstractions—both in his own words as well as chapters written by leading experts in the field, including many of his research collaborators. In addition, this volume contains biographical material, his Turing Award lecture, the transcript of an interview and some of his seminal papers.*

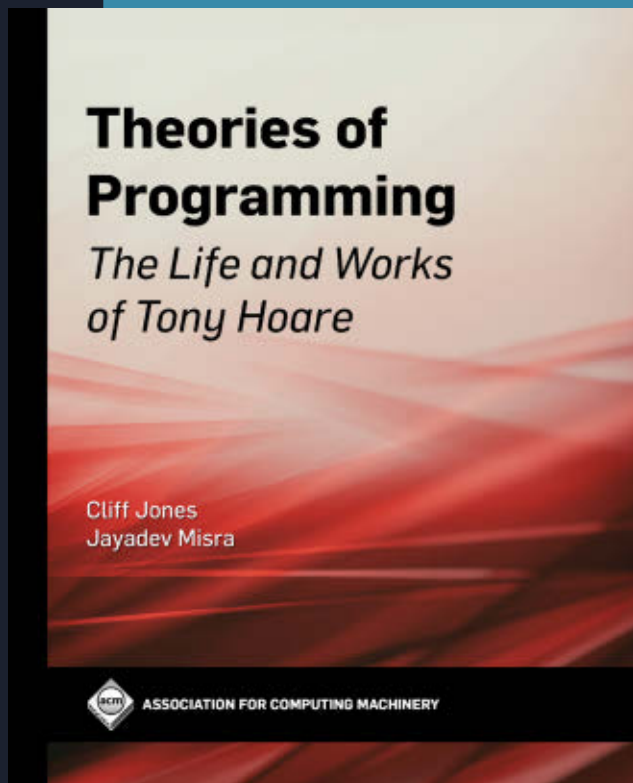
*Hoare's foundational paper "An Axiomatic Basis for Computer Programming," presented his approach, commonly known as Hoare Logic, for proving the correctness of programs by using logical assertions. Hoare Logic and subsequent developments have formed the basis of a wide variety of software verification efforts. Hoare was instrumental in proposing the Verified Software Initiative, a cooperative international project directed at the scientific challenges of large-scale software verification, encompassing theories, tools, and experiments.*

*Tony Hoare's contributions to the theory and practice of concurrent software systems are equally impressive. The process algebra called Communicating Sequential Processes (CSP) has been one of the fundamental paradigms, both as a mathematical theory to reason about concurrent computation as well as the basis for the programming language occam. CSP served as a framework for exploring several ideas in denotational semantics such as powerdomains, as well as notions of abstraction and refinement. It is the basis for a series of industrial-strength tools that have been employed in a wide range of applications.*

*This book also presents Hoare's work in the last few decades. These works include a rigorous approach to specifications in software engineering practice, including procedural and data abstractions, data refinement, and a modular theory of designs. More recently, he has worked with collaborators to develop Unifying Theories of Programming (UTP). Their goal is to identify the common algebraic theories that lie at the core of sequential, concurrent, reactive, and cyber-physical computations.*

<http://books.acm.org>

<http://store.morganclaypool.com/acm>



## Theories of Programming

### *The Life and Works of Tony Hoare*

Edited by

**Cliff Jones**

**Jayadev Misra**

ISBN: 978-1-4503-8729-3

DOI: 10.1145/3477355



# research highlights

---

P. 120

**Technical  
Perspective**  
**Finding the Sweet  
Spot Amid Accuracy  
and Performance**

By Pascal Van Hentenryck

P. 121

## Multi-Itinerary Optimization as Cloud Service

By Alexandru Cristian, Luke Marshall, Mihai Negrea,  
Flavius Stoichescu, Peiwei Cao, and Ishai Menache

---

P. 130

**Technical  
Perspective**  
**On Proofs,  
Entanglement,  
and Games**

By Dorit Aharonov  
and Michael Chapman

P. 131

## $MIP^* = RE$

By Zhengfeng Ji, Anand Natarajan,  
Thomas Vidick, John Wright, and Henry Yuen

# Technical Perspective

## Finding the Sweet Spot Amid Accuracy and Performance

By Pascal Van Hentenryck

THE FIELD OF transportation and logistics has witnessed fundamental transformations in the last decade, due to the convergence of seemingly unrelated technologies. The fast pace of innovations has been particularly striking for an industry that had been relatively stagnant for a long time.

Taxi services were born in England where a public coach service for hire was first documented in 1605. The Hackney Carriage Act, which legalized horse-drawn carriages for hire, was passed in Parliament in 1635, and a similar service was started in Paris in 1637. Public transit was invented by Blaise Pascal in 1662 through a service known as the “carriage,” which was quite popular and operated for 15 years. Both taxi services and public transit adopted new technologies as they became available. Electric battery-powered taxis became available in the streets of London in 1897, and were introduced in New York city the same year. The late 1800s saw the emergence of electric and motor buses. Taxis became widespread in the early 20th century, adopting taxi meters and then, in the late 1940s, two-way radios allowing for communications between drivers and dispatching offices. The automation and optimization of these dispatching services started in the 1980s, but no major evolution took place for several decades thereafter.


This radically changed in the late 2000s through the convergence of multiple technologies, and their embodiment into a single device: the smartphone (the iPhone initially). Transportation network companies (TNCs), such as Uber and Lyft, translated the unique opportunity to connect drivers, riders, and dispatching services everywhere and at massive scale (the “missing ingredient” of Logan Green, co-founder and CEO of Lyft) into novel business models. Ubiquitous connectivity, together with subsequent integrations of GPS navigation, location services, and mapping software, revolutionized transpor-

tation and positioned TNCs as a highly visible face of the digital “gig economy.” GPS-enabled devices also became sensors, collecting the mobility trajectories of millions of users, nowcasting traffic volumes, and estimating travel times. Food, grocery delivery services, as well as crowdsourced “on the way” deliveries for enterprises and small businesses quickly followed. Simultaneously, e-commerce was in the process of fundamentally transforming the shopping experience and the supply chains necessary to sustain it. Packages could now be delivered to front doors, creating massive supply chains and significant challenges in last-mile deliveries.

These last two decades also witnessed impressive progress in optimization technology far from the public view. For instance, mixed integer programming solvers improved by two orders of magnitude from 1998 to 2012, both in terms of speedups in computational times and instances solved within predefined time limits. Optimization solvers were already running significant parts of the economy, dispatching electricity every five minutes to balance generation and consumption, clearing markets for organ exchanges, running steel plants from the furnace to the end products used to build cars, scheduling supply chains, and dispatching logistic systems. But, interestingly, these new economy innovations rely on the ability to connect customers, drivers, and optimization technology through mobile applications and a cloud computing infrastructure.

How this convergence of technologies will impact the economy of the future and society at large is an interesting question to ponder. Will it remain the exclusivity of large corporations and a few startups, or will the software platforms driving this innovation ecosystem become widely available for a wide range of businesses? This open question is precisely why the following paper is exciting: It makes accessible, for the

first time, an end-to-end cloud service that produces traffic-aware, real-time dispatching of agents under complex constraints. The platform leverages GPS traces, traffic predictions, state-of-the-art algorithms for time-dependent shortest paths, large neighborhood search (an optimization technique to find high-quality solutions quickly), and cloud computing to provide multi-itinerary optimization as a service. The authors show that each of these components is critical for the success of the service.

Ignoring traffic conditions (for example, using free-flow speeds) significantly degrades the quality of the service, while optimizing for the worst-case results in largely suboptimal solutions. Similarly, advanced optimization techniques that originated from constraint programming enable the platform to meet the runtime constraints, while capturing the complexity of real-world applications. The paper is particularly timely, partly because of business implications as society may be slowly emerging from a pandemic, and partly because of the agenda it sets for the scientific community. The wide availability of such platforms may be the “missing ingredient” for many businesses to transition to a new economy, democratizing access to technologies that require considerable expertise in many branches of computer science and related disciplines. The paper also highlights the need to model the world with high fidelity in the next generation of optimization algorithms. This is important at a time where society may expect another wave of technology innovations, including drones, autonomous robots, and massive electrification of transportation systems and supply chains. 

**Pascal Van Hentenryck** is the A. Russell Chandler III Chair and a professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech, Atlanta, GA, USA.

Copyright held by author.

# Multi-Itinerary Optimization as Cloud Service

By Alexandru Cristian, Luke Marshall, Mihai Negrea, Flavius Stoichescu, Peiwei Cao, and Ishai Menache

## Abstract

**In this paper, we describe multi-itinerary optimization (MIO)—a novel Bing Maps service that automates the process of building itineraries for multiple agents while optimizing their routes to minimize travel time or distance. MIO can be used by organizations with a fleet of vehicles and drivers, mobile salesforce, or a team of personnel in the field, to maximize workforce efficiency. It supports a variety of constraints, such as service time windows, duration, priority, pickup and delivery dependencies, and vehicle capacity. MIO also considers traffic conditions between locations, resulting in algorithmic challenges at multiple levels (e.g., calculating time-dependent travel-time distance matrices at scale and scheduling services for multiple agents).**

**To support an end-to-end cloud service with turnaround times of a few seconds, our algorithm design targets a sweet spot between accuracy and performance. Toward that end, we build a scalable approach based on the ALNS metaheuristic. Our experiments show that accounting for traffic significantly improves solution quality: MIO finds efficient routes that avoid late arrivals, whereas traffic-agnostic approaches result in a 15% increase in the combined travel time and the lateness of an arrival. Furthermore, our approach generates itineraries with substantially higher quality than a cutting-edge heuristic (LKH), with faster running times for large instances.**

## 1. INTRODUCTION

Route planning and service dispatch operations are a time-consuming manual process for many businesses. This manual process rarely finds efficient solutions, especially ones that must account for traffic, service time windows, and other complicated real-world constraints. Additionally, scale becomes a challenge: service dispatch planning may involve multiple vehicles that need to be routed between numerous locations over periods of multiple days.

The development of large-scale Internet mapping services, such as Google and Bing Maps, creates an opportunity for solving route planning problems automatically, as a cloud service. Large amounts of data regarding geolocations, travel history, etc., are being stored in enterprise clouds and can in principle be exploited for deriving customized itineraries for multiple agents. The goal of such automation is to increase operation efficiency, by determining these itineraries faster (with less man-in-the-loop) and with better quality compared to manually produced schedules. However, multiple challenges must be solved to make this vision a reality.

First, route planning requires efficiently calculating the travel-time matrices between different locations. Although the problem is well understood for free-flow travel times (i.e., no traffic)<sup>13, 8</sup> producing the traffic-aware travel times

on-demand and for any point in time requires careful attention to algorithmic and system scalability. Second, route planning itself must consider multiple features—time windows, priority, amount of time spent in each location (e.g., service duration or dwell time), vehicle capacity, pickup and delivery ordering, and the predicted traffic between locations. The single agent version without all these complex constraints corresponds to the traveling salesman problem (TSP), which is already NP-hard. Numerous extensions to TSP have been studied in operations research and related disciplines under the vehicle routing problem (VRP).<sup>15, 16, 18, 22</sup> However, the bulk of the work is not readily extensible to account for traffic between locations, especially at a large scale. To address customer requirements, our service must incorporate traffic and output an optimized schedule within seconds for instances with hundreds of waypoints.

In this paper, we describe multi-itinerary optimization (MIO), a recently deployed Bing Maps service, available for public use.<sup>2</sup> The design of MIO tackles the above algorithmic challenges, as well as underlying engineering requirements (e.g., efficient use of cloud resources). In particular, our solution consists of a structured pipeline of advanced algorithms. At the lowest layer, we compute travel-time matrices by combining contraction hierarchies (CH)<sup>13</sup> with traffic predictions, resulting in an efficient time-dependent shortest-path algorithm. We then use these matrices for itinerary optimization. Our algorithm for itinerary optimization is built on a popular metaheuristic, adaptive large neighborhood search (ALNS),<sup>21</sup> which pursues an optimal schedule by judiciously choosing between multiple search operators (i.e., repair and destroy). Our search operators have been carefully designed to account for traffic and heterogeneous agents. The entire pipeline is implemented as a cloud service, which is easily accessible through a flexible REST architecture.

We perform extensive evaluations to examine the quality of our end-to-end solution. In particular, we first highlight the significance of accounting for traffic in route planning. Our experiments find that when traffic is ignored during planning, up to 40% of the planned work (e.g., services or dwell time at waypoints) violates its time window constraints when traffic is reintroduced. Furthermore, the combined travel time and lateness of an arrival is 15% higher on average than our traffic-aware approach. On the other hand, using conservative travel times (i.e., the maximum travel

The original version of this paper is entitled “Multi-Itinerary Optimization as Cloud Service (Industrial Paper)” and was published in the *Proceedings of the 27<sup>th</sup> ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

between locations) results in schedules with up to 10% more travel time. Next, we compare MIO to a state-of-the-art heuristic, Lin-Kernighan-Helsgaun (LKH).<sup>15</sup> Our results indicate that MIO obtains significantly higher quality solutions in terms of services satisfied on-time, with less processing time—MIO runs 2× faster than LKH on our publicly released instances that include around 1000 locations.<sup>7</sup>

Although related commercial offerings exist (e.g., see Section 5 for an overview), companies typically obscure some details of the algorithms and/or use some algorithmic components as a black box (e.g., the travel-time calculations). To the best of our knowledge, this paper is the first to report the full algorithmic details of an end-to-end cloud service that produces traffic-aware itineraries for multiple agents. The rest of the paper is organized as follows. Section 2 provides some necessary background; Section 3 outlines the details of our algorithms, as well of our cloud deployment. Section 4 describes our experiments and results. We survey related work in Section 5 and conclude in Section 6.

## 2. BACKGROUND AND MOTIVATION

### 2.1. Internet mapping services

Recent innovations in internet mapping services have created many new business opportunities for large cloud providers such as Microsoft and Google. For example, in the business-to-consumer space, location-based services can be used for more personalized user experience and better targeted advertising, among other things. In the enterprise resource planning context, tasks such as fleet management (e.g., pickup and delivery trucks) and workforce scheduling (e.g., dispatching and routing field technicians) can benefit from automated geographic information system (GIS) services built on accurate and up-to-date geospatial data. However, in the age of the internet, users of mapping services have high expectations. For example, a “large” scheduling task (say, hundreds to thousands of waypoints) is expected to complete within a few minutes, and smaller tasks (tens of waypoints) within seconds. From the cloud provider perspective, these expectations translate to service-level objectives (SLOs) on the end-to-end response time. Additionally, providers wish to generate high-quality solutions on predefined metrics (e.g., minimize the travel time or fuel consumption), while using compute resources efficiently.

### 2.2. Multi-itinerary optimization

Bing Maps recently deployed an enterprise-level planning service called multi-itinerary optimization (MIO).<sup>2</sup> The term “itinerary” corresponds to a single agent (e.g., truck or technician) and has carried over from an earlier consumer version for vacation planning. MIO takes as input a set of waypoints, that is, lat/long locations, each with their own requirements; the requirements may include dwell time, time window, priority, pickup locations, and quantity of goods. The other part of the input is a set of agents. Each agent is characterized by a start/end location, available time window, and vehicle capacity (when relevant). Given this input, the goal of MIO is to find a feasible assignment (i.e., respecting constraints) of a subset of waypoints to the given agents, which maximizes some system objective. For example, a popular objective is to

maximize the number (or priority) of waypoints visited while minimizing the total travel time.

### 2.3. Design challenges

The implementation of MIO as a cloud service has multiple levels of complexity. First and foremost, the service must scale robustly to client demand. Our travel-time calculations involve taking a set of lat/long locations (anywhere in the world), snapping them to the road network, and quickly calculating pairwise shortest travel distance while accounting for traffic. This has required significant algorithmic and engineering effort to support large volumes of users and waypoints. On top of that, incorporating predictive traffic into route planning poses an additional level of complexity to a problem that is already NP-hard. One may wonder whether it is really necessary to account for traffic. Figure 1 demonstrates time-dependent travel times for a single origin-to-destination route (Seattle downtown to Microsoft campus). Notice that the travel time varies considerably throughout the day due to traffic, particularly during peak times (8–9 AM and 3–6 PM). With such significant variations in travel times, it is essential to explicitly account for traffic. Our experiments in Section 4.2 provide a quantitative analysis of this intuition and highlight potential business ramifications when traffic is ignored.

## 3. MIO DESIGN

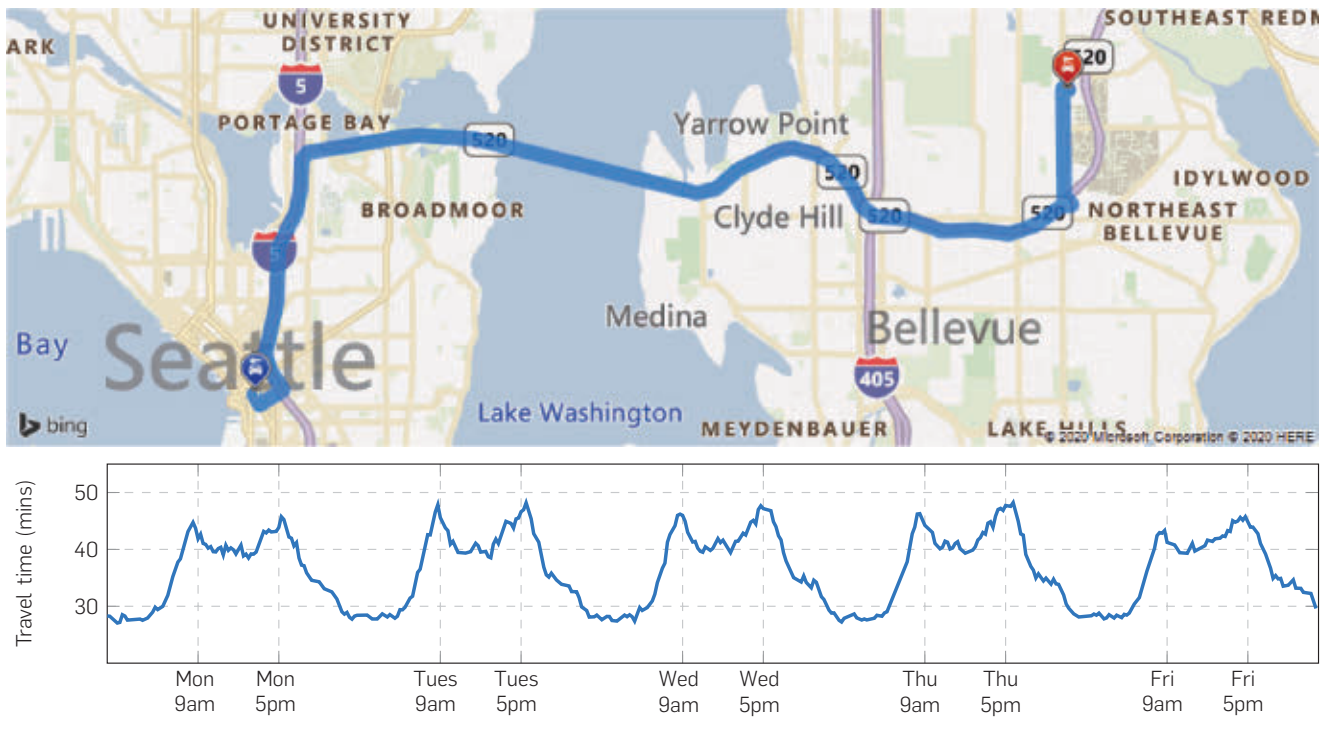
In this section, we provide the details of MIO’s design. Section 3.1 describes how we efficiently calculate travel-time matrices that account for traffic. These travel-time matrices serve as input to our route planning optimization (Section 3.2). In Section 3.3, we highlight some engineering choices that make MIO an efficient cloud service.

### 3.1. Travel-time calculations

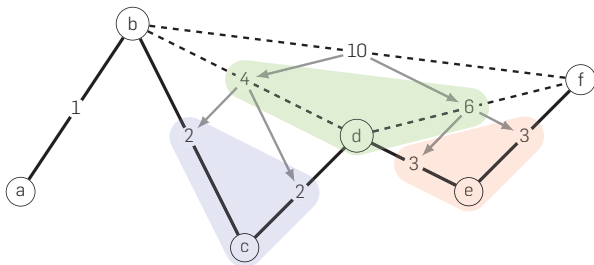
Given a set of locations, a *distance matrix* is a two-dimensional matrix constructed by calculating the length of the shortest path between each pair of locations. The shortest path can in principle represent different metrics, such as physical distance, travel time, and cost. By convention, we use the term distance matrix when the shortest path minimizes free-flow travel time (i.e., no traffic). A *traffic matrix* adds an extra dimension of time, where the shortest path minimizes time-dependent travel over a given time horizon.

For efficient calculation, our algorithm for generating a traffic matrix uses an associated distance matrix as a baseline and extends it to time-dependent travel times using precomputed predictive traffic. There are several implementations for fast distance matrix calculations based on hub labels<sup>8</sup> or contraction hierarchies.<sup>13</sup> Contraction hierarchies is an efficient approach (in data size, preprocessing, and query speed) for distance calculations in road networks. It dramatically reduces the query time required to calculate shortest-path distances by performing a preprocessing step. The preprocessing step generates a multilayered node hierarchy (vertex levels) formed by a ‘contraction’ step, which temporarily removes nodes and adds ‘shortcuts’ to preserve correctness. Figure 2 shows a simple example. Our challenge was to integrate this method of fast computation of travel times, while taking into account traffic fluctuations.

**Figure 1. Time-dependent traffic profile for an example route.**



**Figure 2. An example CH graph.<sup>24</sup> Solid lines represent the original road network. A contraction step temporarily removes nodes (highlighted regions) and adds shortcuts (dashed line). Gray arrows show which edges are combined.**

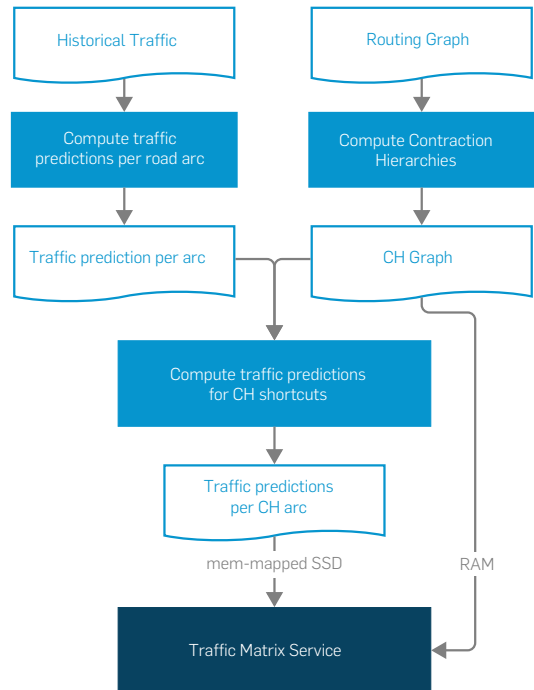


**Input and offline processing.** Our system combines many sources of traffic-related input (e.g., GPS traces) in order to estimate the travel times along every road at any moment in time. To give accurate day-of-week traffic predictions (with 15-minute granularity), we aggregate these travel times for each road, using six months of historical data—giving more weight to recent travel.

After preprocessing our road graph with contraction hierarchies (CH), we compute the travel time for every contraction offline, based on the predictive traffic data and other graph properties such as turn restrictions or turn costs; see Figure 3. As a result of this offline process, we have two outputs:

1. The CH graph (i.e., shortcut graph and vertex levels)
2. The predictive traffic data for each edge in the CH graph containing 672 values (7 days x 24 hours x 4 quarter hours)

**Figure 3. Offline pipeline for traffic matrix service.**



At query time, the CH graph is loaded in memory (~3GB for Western North America; ~85GB for the whole world) and the predictive traffic data is read from SSD using memory-mapped files (~100GB for Western North America; ~5TB for the whole world). The query to calculate the traffic matrix for

a given time horizon uses a time-dependent shortest-path algorithm based on bidirectional Dijkstra.<sup>19</sup>

**Time-dependent shortest path.** In general, the time-dependent shortest-path problem is at least NP-hard; however, under certain assumptions, it can be solved efficiently with polynomial-time algorithms.<sup>9, 14, 11</sup> In particular, we assume the FIFO property, which essentially implies that later departures have later arrivals.

There are several possible objectives for the time-dependent shortest-path problem. For example, if one minimizes travel duration, the solution may wait at the origin until traffic has subsided, whereas when minimizing the total travel time, the solution may wait at any road junction to avoid traffic. In our design, we minimize for arrival time, which avoids waiting and is a common goal in practice. Regardless of the variant, the solution to the time-dependent shortest-path problem is a distance function, parameterized by a starting ‘dispatch’ time. Our algorithm has been specifically designed to produce a (three-dimensional) traffic matrix, that is, a piecewise-constant time-dependent distance function (discretized into 15-minute intervals) for each pair of locations. We base our time-dependent shortest-path algorithm on a bidirectional Dijkstra algorithm, in order to quickly calculate all time-dependent distances from a single origin to many destinations simultaneously.

**Algorithm details.** CH is applied over the road network and the resulting shortcut graph and vertex levels are used. Predicted travel times are stored in 15-minute intervals over one week (for a total of 672 values). The CH graph is loaded into RAM, and traffic predictions are accessed via memory-mapped files on SSD.

A traffic matrix request of  $N$  waypoints over a time horizon with  $T$  intervals performs  $N$  individual one-to- $N$  time-dependent shortest-path queries (one for each origin, for a total of  $N \times N \times T$  travel times). The CH graph is used to expedite the bidirectional (forward/backward) search, as it is sufficient to only explore nodes that have a higher vertex level (i.e., greater importance).

In the forward search, starting from the origin, vertices are explored in the order of their level, by following outgoing hops (edges). For each vertex encountered, the time-dependent travel time is calculated from the origin, and the shortest travel times for each vertex (and requested time intervals) are cached. To improve query speed, the forward search is bounded by the number of hops from the origin ( $2 \times$  maximum number of shortcut edges in a contraction) and distance traveled ( $10 \times$  free-flow travel time from origin to farthest destination).

For each destination, a backward Dijkstra-like search is performed, again by only visiting nodes (via incoming edges) with higher vertex levels in the CH graph and using free-flow travel time to determine the shortest path. Once a vertex is reached by the backward search that has been seen during forward search, the backward route from the found vertex to the destination and the travel times (with predicted traffic) are calculated. Each resulting travel time at each interval is then compared with previously found travel times (initially set to a very high value)—if any are smaller than what has already been found, then the results array is updated. After a fixed maximum number of rounds have been performed

without finding a smaller travel time for any interval, the backward search is stopped.

After all destinations have been processed, the algorithm returns the  $N \times T$  shortest travel times for the given origin node and time horizon.

We note that using free-flow travel times to guide the backward search may result in an approximation to the time-dependent shortest path; however, this is unlikely to occur in practice due to the continued search for better solutions after the first intersection between the forward and backward graphs was found.

**Performance.** We evaluated the traffic matrix performance on a query set of 1200 instances consisting of waypoints from Germany. These queries vary by the distance between waypoints and the matrix request size, with 100 queries per variant. Each query was run with a time horizon of both 96 and 672 intervals (a single day and a full week, respectively), and the output is as shown in Table 1. The average response times were partitioned by distance, number of intervals, and the size of the request matrix—where *cold* indicates the response time when traffic data was read from SSD, and *warm* indicates the response time when the traffic data was already cached in memory.

Observe that distance has little impact on the *warm* response time; however, it significantly impacts the *cold*. This is due to less vertices being shared between the routes, causing more SSD trips to fetch the traffic data, as well as requiring a longer forward exploration phase. On the other hand, the *warm* response time seems to be most impacted by the number of intervals—requesting 672 intervals is clearly more expensive than 96 intervals for both *cold* and *warm*, but the *warm* response time for the full week is almost equivalent to the *cold*. Finally, changes in matrix size impacts performance as expected—all experiments were run single threaded, hence the linear relationship between the single source matrix (e.g.,  $1 \times 10$ ) and the multiple source matrix (e.g.,  $10 \times 10$ ); this is most noticeable for the *warm* response time (e.g., 0.03 and 0.3 s, respectively).

**Table 1. Traffic matrix cold/warm response times.**

Size	Intervals	Distance (km)	Cold (s)	Warm (s)
10×10	96	<15	0.1224	0.0656
10×10	96	<80	0.3620	0.0810
10×10	96	<250	1.0522	0.0887
10×10	96	<15	0.1224	0.0656
10×10	672	<15	0.3384	0.3085
1×10	672	<15	0.0786	0.0363
10×10	672	<15	0.3384	0.3085
1×100	672	<15	0.2163	0.1737
100×100	672	<15	17.8519	17.8524

We compare the effect of distance between locations, number of intervals, and size of request.

### 3.2. Itinerary optimization

Equipped with traffic matrices, MIO finds efficient solutions to various different routing scenarios and objectives. At its core, it schedules agents to visit a set of locations. Each location is visited at most once, by at most one agent, unless the location is a depot or supports pickups/dropoffs. The dwell time at each location is given and must be completed within the specified time window at that location. Agents may arrive early, but they must wait until the time window before starting their service. Vehicles may have a limited capacity when delivering goods from a depot, or moving items from pickup locations. It is currently assumed that items cannot be split and cannot be transferred to other agents or routes by an intermediate location. Our default objective is to maximize the number of visited locations, weighted by priority, while minimizing the total travel time of the agents.

**MIO algorithmic approach.** Adaptive large neighborhood search (ALNS) is a popular metaheuristic that has been used for many optimization problems and in particular for vehicle routing problems. ALNS uses a simulated annealing framework, with a local search at each iteration that adapts its behavior based on previous iterations. This local search is controlled by randomly choosing a pair of destroy/repair operations (from a predefined set). Then, starting from a feasible solution, the destroy “local search” modifies the solution, such that the solution may become infeasible. The repair operation then alters the solution with a guarantee of feasibility. If the new solution is better than the previous, then the probability of choosing these destroy/repair operations will increase. This is the adaptive learning component of ALNS. Our implementation follows a similar approach as outlined in Pisinger and Ropke,<sup>21</sup> and in addition, we support parallel processing and multiple objectives (via an automatic weighting scheme). See Algorithm 1 for a high-level overview of the approach. The termination condition for ALNS is met when either the incumbent solution has not changed within the last 5000 iterations after we reach a minimum temperature (i.e., we are confident that we explored enough of the search space and the solution did not change) or a fixed timeout, given by a function of agents and locations, has been reached. At the end of this process, we apply a k-opt swap post-optimization step on our solution, in a final attempt to improve its quality.

The most important component for efficient computation is the design of the destroy and repair operators. These guide the local search and attempt to exploit the structure of the combinatorial optimization problem. A careful balance must be considered—a trade-off between fast iterations and intelligent decisions. Here, we list our set of operators, with a brief description of their intent.

**Destroy methods** **SKIPS**: removes visits where the cost of reaching the next location is greater than going directly from the previous one; **HIGH COST**: removes the most ‘expensive’ (e.g., in terms of travel distance) locations; **REPLACE ITEM**: removes a location and replaces it with another suitable one; **NEIGHBOURHOOD**: picks a random location and then removes the locations in its neighborhood; **TRANSFERS**: removes locations that would have a lower cost in the itinerary

of other agents.

**Repair methods** **EARLY**: prefers to insert random locations near the beginning of an agent’s schedule; **LATE**: prefers to insert random locations near the end of an agent’s schedule; **NN**: inserts locations using the nearest neighbor heuristic; **greedy**: inserts locations using a **GREEDY** algorithm.

---

#### Algorithm 1: High-level ALNS algorithm

---

```
// x: Initial feasible solution
// T: Initial temperature
//  $\alpha$ : Cooling rate
//  $\pi$ : Initial adaptive weights
1 def ALNS_SOLVE( $x, T, \alpha, \pi$ ):
2    $X_{best} \leftarrow x$ 
3
4   while not terminated do
5     REPAIR, DESTROY  $\leftarrow$  CHOOSE_PAIR( $\pi$ )
6      $x' \leftarrow$  REPAIR(DESTROY( $x$ ))
7
8     if OBJ( $x'$ ) > OBJ( $X_{best}$ ) then
9        $result \leftarrow$  NewIncumbent
10       $X_{best} \leftarrow x'$ 
11       $x \leftarrow x'$ 
12     else if OBJ( $x'$ ) > OBJ( $x$ ) then
13        $result \leftarrow$  DominatesCurrent
14        $x \leftarrow x'$ 
15     else if SIMULATED_ANNEALING( $x, x', T$ ) then
16        $result \leftarrow$  Accepted
17        $x \leftarrow x'$ 
18     else
19        $result \leftarrow$  Rejected
20     end
21
22      $\pi \leftarrow$  UPDATE( $\pi, result, REPAIR, DESTROY$ )
23      $T \leftarrow \alpha T$ 
24   end
25 return  $X_{best}$ 
```

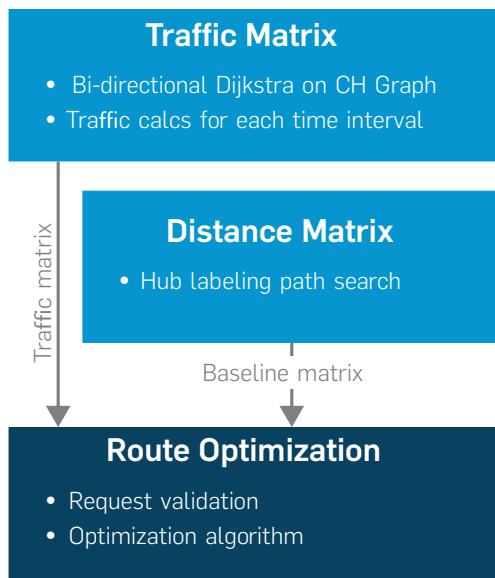
---

### 3.3. Cloud deployment and engineering

The design of MIO as a cloud service must consider hosting costs and resource usage. Large amounts of data must be stored in memory for quick access (hub labels and the CH graph  $\sim 200$ GB) and on SSD (traffic data  $\sim 5$ TB), and many underlying virtual machines (VMs) may be required to support a large volume of requests. A simple implementation would consist of a single VM role that hosts the entire service, where the number of the VMs can be scaled based on request volume. However, we can exploit the special structure of our service for a more resource-efficient architecture. From an operational perspective, we have three different components that support the entirety of the algorithms described earlier in this section:

1. *Distance Matrix* requires a lot of memory to host the hub labels.
2. *Traffic Matrix* needs little memory to host the CH graph. However, the component is CPU intensive and

**Figure 4. System architecture. Route optimization can take either traffic or distance matrix as input.**



requires a lot of SSD storage for the traffic data.

3. *Route Optimization* does not persist any data in memory; the computation itself is CPU intensive.

Accordingly, the MIO system includes three different VM roles, one for each of the above components (see Figure 4 for a high-level architecture overview). The roles communicate via binary HTTP protocol. Each component can be separately scaled up/down based on current load conditions. Specifically, the distance matrix role is hosted on memory-optimized VMs<sup>1</sup>; because our design achieves high throughput, we rarely need to scale this component. For the traffic matrix, SSD storage needs to be provisioned for the maximum load conditions. In addition, we use compute-optimized VMs<sup>1</sup> for its computation; these VMs can be scaled quickly to reduce operating costs. Similarly, the route optimization role is also hosted on compute-optimized VMs that are scaled as necessary. Clearly, the alternative implementation with a single VM role would not be able to achieve this level of flexibility and efficiency—any scale up is likely to result in resource wastage in at least one dimension.

To enable the above, each role is composed of:

1. A dynamic scale set of VMs that scales based on a custom performance counter; the counter is tuned to the needs of the specific role.
2. An Azure load balancer that spreads the requests to the VM scale set.
3. An Azure traffic manager that handles the distribution of requests to the closest datacenter, as well as failover management in case of outage.

We conclude this section by briefly describing the interface of MIO. Some applications need to calculate small schedules with very little latency. Although other applications have larger problems and more lenient time frames,

they would like to run the service in the background while displaying perhaps a progress bar for the optimization. To address these different needs, we have designed MIO with both synchronous and asynchronous interfaces (i.e., for the former and latter applications, respectively). The synchronous interface receives an optimization request and responds with an optimized itinerary, whereas the asynchronous interface returns a callback id and schedules the job to be executed in the background. An Azure queue is used to orchestrate the execution of the background job on the service side, and the resulting itinerary is saved in Azure storage. When the client uses the callback id to poll for the completion of the job, MIO checks if the optimization result exists in Azure storage and, if so, sends the output back to the application.

## 4. EXPERIMENTAL RESULTS

### 4.1. Experiment setup

To evaluate the impact of traffic and the performance of MIO compared to state-of-the-art heuristics, we performed a large-scale computational study. In particular, we use the publicly available instances<sup>7</sup> defined in a recent Microsoft Research project<sup>17</sup>. These instances have been designed to realistically model the technician routing and scheduling problem with uncertain service duration. The 1440 instances use real-world locations and include a simplified traffic matrix obtained from our service. They cover a range of different sizes and have multiple agents with various shift starting times.

For a fair comparison of our algorithms, we use the expected value of the service duration and ignore instances where it is impossible to visit all waypoints when using maximum travel times. Accordingly, our optimization goal is to visit all waypoints while minimizing the travel time. If an algorithm cannot guarantee that time windows are satisfied (i.e., the agent is late to a waypoint), we prioritize minimizing lateness over minimizing travel time.

We compare traffic-agnostic and traffic-aware variants of MIO to heuristics based on Lin-Kernighan. These heuristics have been specifically designed to quickly find high-quality solutions to symmetric TSPs. For our experiments, we use LKH v3.0.6,<sup>15</sup> a state-of-the-art extension to Lin-Kernighan that supports constrained optimization of asymmetric TSPs. LKH has been designed to optimize many different VRP scenarios and has had great success at incredibly large-scale instances.

We configured LKH as a capacitated vehicle routing problem with time windows (CVRPTW), which also supports dwell times at locations. Demands were set to zero. The maximum travel time between pairwise locations was used to avoid time-window penalties when calculating the true cost with traffic. Default parameters were used, with the addition of “SPECIAL”—we found that excluding this parameter increased the computation time for some instances by over 1500×, with little difference in solution quality.

All experiments ran on an Azure Standard DS14v2 VM instance having 16 cores and 120GB RAM. Although MIO can exploit multiple cores, we restrict all algorithms



to a single thread for a fair comparison. Experiments of the same algorithm were run in parallel on the same VM instance (i.e., up to 16 experiments running concurrently) for faster throughput.

## 4.2. Experiment design

In our experiments, we wish to examine the significance of incorporating traffic. To that end, we run MIO in its standard mode with traffic predictions (MIO) and compare the results to settings where the traffic predictions are replaced with static travel times. In particular, we execute MIO under two variants of traffic-agnostic settings:

- MIN: minimum time between locations (i.e., traffic-free)
- MAX: maximum time between locations (i.e., peak traffic over a one-week period)

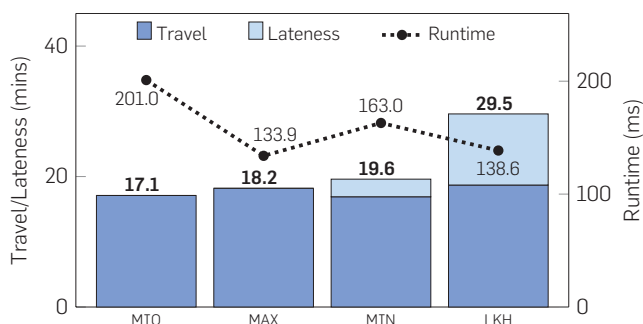
We note that using minimum travel, as in MIN, has historically been the most common approach, as free-flow travel times can be accurately estimated using only distance and the enforced speed limits—instead of requiring the large volumes of traffic data (see Section 3.1.1). As discussed above, we compare the different MIO variants with the LKH algorithm, using maximum travel between locations (LKH).

By the design of our experiments, it is possible to visit all waypoints within their respective time windows; however, the traffic-agnostic solutions may result in scheduling a waypoint that violates its time window. That is, the agent would arrive late when reintroducing the actual travel time (capturing the predicted traffic). When such a violation occurs, we record the lateness of the agent and increase the count of violations. More specifically, MAX and MIO do not violate any time windows (i.e., the agent is never late), whereas MIN and LKH might. MIN is expected to violate time windows, as it is unaware of extreme traffic conditions. LKH seems to treat time windows as soft constraints; thus, some lateness is expected, although we configure the algorithm to prioritize arriving on time.

## 4.3. Results

A summary of our results is given in Figure 5; breakdowns by instance size are provided in Table 2 and Figure 6. All results have been normalized by the number of waypoints in an instance, that is, they represent average cost per waypoint.

**Figure 5. Comparison of algorithms over the average travel time and lateness for each waypoint visited.**

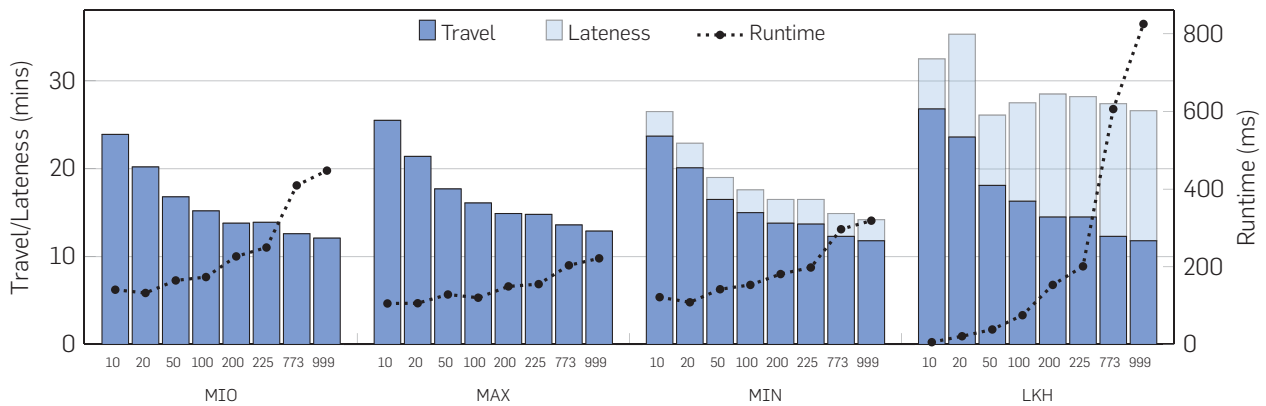


For example, Figure 5 implies that for every additional waypoint MIO requires an additional 17.1 minute travel time and takes an extra 201 ms runtime to solve (on average). This allows a fair comparison of instances with different sizes. In particular, Figure 6 highlights the “economy of scale” of large instances using the same geographic area—travel cost per waypoint decreases with the number of waypoints, which is expected as agents need to travel less between locations.

As can be seen from the results, MAX requires over 6% more travel per waypoint than MIO, which incorporates traffic. Nonetheless, the average runtime of MAX was the fastest of all algorithms. This is due to the ultra quick cache hit of static travel times; that is, MAX does not need a binary search for time-of-day lookup; in addition, using maximum travel leads to fewer valid routes for the algorithm to evaluate. In practice, using maximum travel time is not recommended with a hard time-window constraint, as waypoints might never be scheduled, because they “appear” to be impossible to reach within the time window. In contrast, MIN unknowingly compromises travel time with lateness. Consequently, its travel time is 1% less than MIO on average, but its combined travel and lateness is almost 15% larger.

**Table 2. Comparison of results per waypoint visited.**

#WOs/agents	Algo.	Travel (mins)	Lateness (mins)	Runtime (ms)	#Violations (count)
10/3 [228]	MIO	23.9	0.0	141.0	0.00
	MAX	25.5	0.0	105.6	0.00
	MIN	23.7	2.8	121.8	0.15
	LKH	26.8	5.7	5.9	0.07
20/7 [229]	MIO	20.2	0.0	132.6	0.00
	MAX	21.4	0.0	106.2	0.00
	MIN	20.1	2.8	108.6	0.16
	LKH	23.6	11.7	21.2	0.13
50/15 [135]	MIO	16.8	0.0	164.9	0.00
	MAX	17.7	0.0	128.7	0.00
	MIN	16.5	2.5	142.2	0.15
	LKH	18.1	8.0	38.6	0.10
100/30 [233]	MIO	15.2	0.0	173.6	0.00
	MAX	16.1	0.0	120.4	0.00
	MIN	15.0	2.6	153.4	0.15
	LKH	16.3	11.2	75.4	0.14
200/60 [231]	MIO	13.8	0.0	226.8	0.00
	MAX	14.9	0.0	149.5	0.00
	MIN	13.8	2.7	181.3	0.16
	LKH	14.5	14.0	153.4	0.17
225/75 [78]	MIO	13.9	0.0	249.7	0.00
	MAX	14.8	0.0	155.4	0.00
	MIN	13.7	2.8	198.2	0.16
	LKH	14.5	13.7	201.2	0.17
773/275 [79]	MIO	12.6	0.0	409.6	0.00
	MAX	13.6	0.0	203.8	0.00
	MIN	12.3	2.6	296.7	0.15
	LKH	12.3	15.1	606.2	0.18
999/350 [73]	MIO	12.1	0.0	447.7	0.00
	MAX	12.9	0.0	221.8	0.00
	MIN	11.8	2.4	318.8	0.15
	LKH	11.8	14.8	825.3	0.18

**Figure 6. Comparison of algorithms, with results averaged over each waypoint visited.**

Finally, we compare our results with LKH. For smaller instances, LKH is extraordinarily fast, being, on average, over  $17\times$  faster than our MIO variants—but this changes at scale, with MIO being almost  $2\times$  faster than LKH. Unfortunately, LKH seems to struggle finding high-quality solutions across the board. As LKH uses maximum travel times, we expect the results to be similar to MAX; however, LKH seems to generate schedules with significantly larger lateness. The combined travel and lateness is over  $72\%$  higher than MIO, that is, almost  $2\times$  greater on average. Careful inspection of our configuration and the publicly available LKH code suggests that this particular issue with low-quality solutions is systemic to the LKH algorithm. Although we could not identify anything incorrect with their approach, we did not conduct a full analysis to see if the issues are inherent to the Lin-Kernighan approach, or within the extensions added by LKH.

## 5. RELATED WORK

Most literature relating to vehicle routing problems (VRP) makes the assumption that the travel time between locations is given as input.<sup>23</sup> Often, these problems have a fixed set of locations (i.e., the distance matrix can be cached), but in general, the effort to optimize the VRP is significantly complex so that the shortest-path travel-time calculations are typically ignored. We, in contrast, pay close attention not only to the route optimization but also to the efficient calculation of travel times. Our system provides a complete end-to-end service that supports waypoints anywhere across the globe and aims to give a high-quality solution within a very short time frame. In our service, we have implemented the necessary distance and traffic calculations between locations, which are algorithmically nontrivial.

### 5.1. Related commercial products

Enterprises offer related cloud services, such as Microsoft Dynamics 365 for Field Service (Resource Scheduling Optimization),<sup>4</sup> Routific Routing Engine,<sup>5</sup> Google Maps Directions,<sup>3</sup> and TomTom Routing.<sup>6</sup> Such commercial offerings typically do not disclose the full details of the algorithms and/or use some algorithmic components as black box (e.g., the travel-time calculations). To the best of our knowledge, this paper is the first to report the full algorithmic details of

an end-to-end cloud service that supports traffic-aware itineraries for multiple agents.

### 5.2. Time-dependent shortest path

For an overview of query acceleration techniques for time-dependent shortest-path algorithms, we refer to Delling and Wagner.<sup>10</sup> An approach similar to our traffic matrix algorithm was described in Geisberger<sup>12</sup>, who also uses contraction hierarchies with a modified Dijkstra algorithm to calculate time-dependent shortest travel between a single origin and multiple destinations.

### 5.3. The vehicle routing problem (VRP)

MIO was originally designed to solve the technician routing and scheduling problem (TRSP), which is a variant of VRP. The main extension of TRSP over VRP is the dwell time at each location, which can be different for each agent (i.e., based on skill/experience). Since our first design, MIO has been extended to include additional VRP features, such as vehicle capacity, pickup and delivery, multiple depots, and more. The ALNS method was originally introduced in Pisinger and Ropke<sup>21</sup> as a general-purpose approach for solving VRPs and has been used in numerous papers, for example, to solve the TRSP.<sup>16,20</sup>

## 6. CONCLUSION


In this paper, we described multi-itinerary optimization (MIO)—a Bing Maps service that is publicly available worldwide. MIO takes a list of agents and desired waypoints with various requirements and outputs an efficient schedule and route for each agent. MIO encompasses a variety of algorithms, such as shortest-path mechanisms for travel-time calculation and a carefully designed heuristic for route optimization. Importantly, our algorithms account for *traffic* predictions, which have significant impact in urban areas. Our experiments show that MIO achieves efficient solutions with fast performance—it produces high-quality schedules at scale with running times better than a state-of-the-art heuristic (LKH) approach. We see MIO as an attractive tool that can be used to automate relevant logistics operations of numerous businesses and organizations. G

## References

- Azure pricing. <https://azure.microsoft.com/en-us/pricing/details/virtual-machines/windows>.
- Bing Maps Multi-Itinerary Optimization. <https://microsoft.com/en-us/maps/fleet-management>.
- Google Maps Routes. <https://cloud.google.com/maps-platform/routes>.
- Resource Scheduling Optimization (RSO). <https://dynamics.microsoft.com/en-us/field-service>.
- Routific. <https://routific.com>.
- TomTom Routing API. <https://developer.tomtom.com/routing-api>.
- TRSP instances. <https://github.com/microsoft/trsp>.
- Abraham, I., Dellling, D., Goldberg, A.V., Werneck, R.F. A hub-based labeling algorithm for shortest paths in road networks. In *Experimental Algorithms*, P. M. Pardalos and S. Rebennack, eds. *Lecture Notes in Computer Science* (2011), Springer, Berlin Heidelberg, 230–241.
- Dean, B.C. Shortest paths in FIFO time-dependent networks: Theory and algorithms. *Rapport Technique*, 2004, 13.
- Delling, D., Wagner, D. Time-dependent route planning. In *Robust and Online Large-Scale Optimization: Models and Techniques for Transportation Systems*, R. K. Ahuja, R. H. Möhring, and C. D. Zaroliagis, eds. *Lecture Notes in Computer Science* (2009), Springer, Berlin, Heidelberg, 207–230.
- Foschini, L., Hershberger, J., Suri, S. On the complexity of time-dependent shortest paths. *Algorithmica* 4, 68 (2014), 1075–1097.
- Geisberger, R. Engineering time-dependent one-to-all computation. *arXiv:1010.0809 [cs]* (2010).
- Geisberger, R., Sanders, P., Schultes, D., Dellling, D. Contraction hierarchies: Faster and simpler hierarchical routing in road networks. In *Experimental Algorithms*, C. C. McGeoch, ed. *Volume 5038* (2008), Springer, Berlin, Heidelberg, 319–333.
- He, E., Boland, N., Nemhauser, G., Savelsbergh, M. Computational complexity of time-dependent shortest path problems. *Optimization Online* (2019), 12.
- Helsgaun, K. *An Extension of the Lin-Kernighan-Helsgaun TSP Solver for Constrained Traveling Salesman and Vehicle Routing Problems*. (2017).
- Kovacs, A.A., Parragh, S.N., Doerner, K.F., Hartl, R.F. Adaptive large neighborhood search for service technician routing and scheduling problems. *J. Schedul.* 5, 15 (2012), 579–600.
- Marshall, L., Tankayev, T. Practical risk modeling for the stochastic technician routing and scheduling problem. *Optim. Online* (2020).
- Miranda, D.M., Conceição, S.V. The vehicle routing problem with hard time windows and stochastic travel and service time. *Exp. Syst. Appl.*, 64 (2016), 104–116.
- Nicholson, T.A.J. Finding the shortest route between two points in a network. *Comput. J.* 3, 9 (1966), 275–280.
- Pillac, V., Guéret, C., Medaglia, A.L. A parallel matheuristic for the technician routing and scheduling problem. *Optim. Lett.* 7, 7 (2013), 1525–1535.
- Pisinger, D., Ropke, S. A general heuristic for vehicle routing problems. *Comput. Oper. Res.* 8, 34 (2007), 2403–2435.
- Taş, D., Dellaert, N., van Woensel, T., de Kok, T. Vehicle routing problem with stochastic travel times including soft time windows and service costs. *Comput. Oper. Res.* 1, 40 (2013), 214–224.
- Ticha, H.B., Absi, N., Feillet, D., Quilliot, A. Vehicle routing problems with road-network information: State of the art. *Networks* 3, 72 (2018), 393–406.
- Wikipedia contributors. Contraction hierarchies—Wikipedia, the free encyclopedia, 2020. Online [Accessed: 17-March-2020].

**Alexandru Cristian, Luke Marshall, Mihai Negrea, Flavius Stoichescu, Peiwei Cao, Ishai Menache** ([alexandru.cristian, luke.marshall, mihai.negrea, flavius.stoichescu, peiweic, ishai]@microsoft.com), Microsoft Research, Redmond, WA, USA, and Timiș, România.


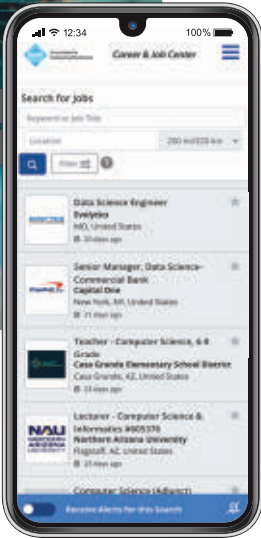
© 2021 ACM 0001-0782/21/11 \$15.00



Association for Computing Machinery

**Career & Job Center**




## The #1 Career Destination to Find Computing Jobs.

**Connecting you with top industry employers.**

**Your next job is right at your fingertips. Get started today!**

### The new ACM Career & Job Center offers job seekers a host of career-enhancing benefits, including:

-  Access to new and exclusive career resources, articles, job searching tips and tools.
-  Gain insights and detailed data on the computing industry, including salary, job outlook, 'day in the life' videos, education, and more with our new Career Insights.
-  Redesigned job search page allows you to view jobs with improved search filtering such as salary, location radius searching and more without ever having to leave the search results.
-  Receive the latest jobs delivered straight to your inbox with **new exclusive Job Flash™ emails**.
-  Get a free resume review from an expert writer listing your strengths, weaknesses, and suggestions to give you the best chance of landing an interview.
-  Receive an alert every time a job becomes available that matches your personal profile, skills, interests, and preferred location(s).

**Visit <https://jobs.acm.org/>**

# Technical Perspective

## On Proofs, Entanglement, and Games

By Dorit Aharonov and Michael Chapman

**WHAT IS A PROOF?** Philosophers and mathematicians have pondered this question for centuries. Theoretical computer science offers a rigorous handle on this deep question. One can think of a proof as a two-player game: an all-powerful though untrusted prover who provides a proof of the statement, and a computationally weak verifier who needs only to verify it. In fact, NP problems can be presented exactly in this verifier-prover language. Viewing proofs as games turned out to be remarkably fruitful. For example, interactive proofs were invented, resembling Socratic dialogues; these are games in which the prover and verifier exchange (possibly randomized) messages. And, why just one prover? In multi-prover interactive proofs (MIP) several non-communicating provers are involved. This gave birth to beautiful concepts such as zero knowledge and probabilistically checkable proofs (PCPs) with immense impact not only theoretically but also in practice, for example, in digital currency.

The following paper studies quantum interactive proofs. Here the provers are allowed to share an entangled quantum state; this resembles sharing a random bit string, except quantum states have those funny, stronger-than-classical correlations; a prototypical example is the Einstein-Podolsky-Rosen (EPR) state, which was said by Einstein to allow “spooky action at a distance.” Can quantum correlations be used to prove stronger statements?

The class of problems solvable by multi-prover quantum interactive proofs is denoted MIP\*. Its computational power had been open for over 15 years. In the following tour-de-force paper, the culmination of a remarkable line of works, the question is finally settled. While it was believed at first that MIP\* equals its classical counterpart MIP, it turns out that

MIP\* contains the halting problem! The result not only resolved the stubborn MIP\* question, but at the same time solved several major decades-old open problems in mathematics. What ideas make this result so far reaching?

Best to start our journey into the MIP\* rabbit hole with the landmark 1991 result, showing that MIP equals NEXP, a scaled-up version of NP where the verifier needs to check exponentially long proofs. In MIP, the verifier only exchanges polynomially many bits with the provers—How can it verify the proof without fully reading it? The key tool is low degree tests, which leverage robustness of low-degree polynomials to allow the verifier to check, using little communication, that both provers hold consistent such polynomials. In 2011, it was shown that the protocol is secure even against entangled provers. MIP\* is thus at least as powerful as its classical counterpart.


But the point of the current paper is that MIP\* is provably much larger than MIP. How can entanglement help? Insight can be drawn from a 2019 paper, which showed for the first time that MIP\* is strictly larger than MIP. The paper proved that MIP\* contains the class of problems requiring doubly exponential long proofs (NEEXP). How can the verifier check such a proof, when even specifying a single location in the proof requires exponentially many bits?

Entanglement comes to the rescue. More precisely, a remarkable property of entangled states called self-testing. In 1964, Bell had devised a game in which two players who share an EPR state can win with a strictly larger probability than classically possible. Uniqueness of the players’ optimal strategy implies that achieving maximal success probability serves as a certificate that their state and measurements are the unique optimal ones. This is a remarkable

form of rigidity of quantum mechanics. It is as if the verifier gets to peek into the secret quantum labs of the provers! Moreover, a quantum version of low-degree tests enables self-tests that verify exponentially long entangled states, using only polynomial communication. Now comes a mind-boggling idea: though the verifier’s messages are too short to specify its exponentially long questions about the doubly exponentially long proof, it can use quantum self-testing to efficiently force the provers to share an exponentially large, entangled state and then to correctly sample their questions themselves.

This result is just the start of the story; going all the way to the halting problem requires modifying the “compression-by-entanglement” idea so it can be applied recursively. A plethora of new hurdles arise, whose solution is the tour de force in this paper.

The exciting applications include constructing infinite algebras that cannot be approximated in any finite dimension and separating quantum correlations models previously conjectured to be equal. Self-testing is connected to group stability, which raises hope for progress on major problems in group theory.

How to explain the strong impact this theoretical computer science result has on pure mathematics? PCPs and other powerful computational complexity concepts are applied here, but perhaps another aspect is the role protocols play in the result. These sequences of individual steps that depend on time, constitute an intuitive way to think about highly complex mathematical objects, an approach that seems to offer a fresh look at physics and mathematical problems. 

Dorit Aharonov is a professor and Michael Chapman is a Ph.D. student in the CS department at the Hebrew University of Jerusalem, Israel.

Copyright held by authors.

# MIP\* = RE

By Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen

**Note from the Research Highlights Co-Chairs:**  
 A Research Highlights paper appearing in *Communications* is usually peer-reviewed prior to publication. The following paper is unusual in that it is still under review. However, the result has generated enormous excitement in the research community, and came strongly nominated by SIGACT, a nomination seconded by external reviewers.

## Abstract

The complexity class NP characterizes the collection of computational problems that have *efficiently verifiable solutions*. With the goal of classifying computational problems that seem to lie beyond NP, starting in the 1980s complexity theorists have considered extensions of the notion of efficient verification that allow for the use of *randomness* (the class MA), *interaction* (the class IP), and the possibility to interact with *multiple proofs, or provers* (the class MIP). The study of these extensions led to the celebrated PCP theorem and its applications to hardness of approximation and the design of cryptographic protocols.

In this work, we study a fourth modification to the notion of efficient verification that originates in the study of *quantum entanglement*. We prove the surprising result that every problem that is recursively enumerable, including the Halting problem, can be efficiently verified by a classical probabilistic polynomial-time verifier interacting with two all-powerful but noncommunicating provers sharing entanglement. The result resolves long-standing open problems in the foundations of quantum mechanics (Tsirelson’s problem) and operator algebras (Connes’ embedding problem).

## 1. INTERACTIVE PROOF SYSTEMS

An *interactive proof system* is an abstraction that generalizes the intuitively familiar notion of *proof*. Given a formal statement  $z$  (e.g., “this graph admits a proper 3-coloring”), a proof  $\pi$  for  $z$  is information that enables one to check the validity of  $z$  more efficiently than without access to the proof (in this example,  $\pi$  could be an explicit assignment of colors to each vertex of the graph, which is generally easier to verify than to find).

Complexity theory formalizes the notion of proof in a way that emphasizes the role played by the verification procedure. To explain this, first recall that a *language*  $L$  is identified with a subset of  $\{0, 1\}^*$ , the set of all bit strings of any length, that represents all problem instances to which the answer should be “yes.” For example, the language  $L = 3\text{-COLORING}$  contains all strings  $z$  such that  $z$  is the description (according to some prespecified encoding scheme) of a 3-colorable graph  $G$ . We say that a language  $L$  admits efficiently verifiable

proofs if there exists an algorithm  $V$  (formally, a polynomial-time Turing machine) that satisfies the following two properties: (i) for any  $z \in L$  there is a string  $\pi$  such that  $V(z, \pi)$  returns 1 (we say that  $V$  “accepts” the proof  $\pi$  for input  $z$ ) and (ii) for any  $z \notin L$ , there is no string  $\pi$  such that  $V(z, \pi)$  accepts. Property (i) is generally referred to as the *completeness* property and (ii) is the *soundness*. The set of all languages  $L$  such that there exists a verifier satisfying both completeness and soundness is the class NP.

Research in complexity and cryptography in the 1980s and 1990s led to a significant generalization of this notion of “efficiently verifiable proof.” The first modification is to allow *randomized* verification procedures by relaxing (i) and (ii) to *high probability* statements: every  $z \in L$  should have a proof  $\pi$  that is accepted *with probability at least  $c$*  (the completeness parameter), and for no  $z \notin L$  should there be a proof  $\pi$  that is accepted *with probability larger than  $s$*  (the soundness parameter). A common setting is to take  $c = \frac{2}{3}$  and  $s = \frac{1}{3}$ ; standard amplification techniques reveal that the exact values do not significantly affect the class of languages that admit such proofs, provided that they are chosen within reasonable bounds.

The second modification is to allow *interactive* verification. Informally, instead of receiving a proof string  $\pi$  in its entirety and making a decision based on it, the verification algorithm (called the “verifier”) now communicates with another algorithm called a “prover,” and based on the interaction decides whether  $z \in L$ . There are no restrictions on the computational power of the prover, whereas the verifier is required to run in polynomial time.<sup>a</sup> We let IP (for “Interactive Proofs”) denote the class of languages having interactive, randomized polynomial-time verification procedures.

The third and final modification is to consider interactions with *two* (or more) provers. In this setting, the provers are not allowed to communicate with each other and the verifier may “cross-interrogate” them in order to decide if  $z \in L$ . The provers are allowed to coordinate a joint strategy ahead of time, but once the protocol begins, they can only interact with the verifier. The condition that the provers cannot communicate with each other is a powerful constraint that can be leveraged by the verifier to detect attempts at making it accept a false claim, that is whenever  $z \notin L$ .

<sup>a</sup> The reader may find the following mental model useful: in an interactive proof, an all-powerful prover is trying to convince a skeptical, but computationally limited, verifier that a string  $z$  (known to both) lies in the set  $L$ , even when it may be that in fact  $z \notin L$ . By interactively interrogating the prover, the verifier can reject false claims, i.e. determine with high statistical confidence whether  $z \in L$  or not. Importantly, the verifier is allowed to probabilistically and adaptively choose its messages to the prover.

The original version of this paper appeared on the quant-ph arXiv as arXiv:2001.04383.

Work in the 1980s and 1990s in complexity theory has shown that the combination of randomness, interaction, and multiple provers leads to an *exponential* increase in verification power. Concretely, the class of problems that can be verified with all three ingredients together, denoted MIP (for Multiprover Interactive Proofs), was shown to equal the class NEXP<sup>1</sup> of languages that admit exponentially long “traditional” proofs verifiable in exponential time.

## 2. OUR RESULTS

We now introduce a fourth modification to the class MIP, leading to the class MIP\* that is the focus of our work. Informally the class MIP\* contains all languages that can be decided by a classical polynomial-time verifier interacting with multiple *quantum* provers sharing *entanglement*. To be clear: compared with MIP, the only difference is that the provers may use entanglement, both in the case when  $z \in L$  (completeness) and when  $z \notin L$  (soundness).

The study of MIP\* is motivated by a long line of works in the foundations of quantum mechanics around the topic of *Bell inequalities*. Informally, a Bell inequality is a linear function that separates two convex sets: on the one hand, the convex set of all families of distributions that can be generated locally by two isolated parties using shared randomness; on the other hand, the convex set of all families of distributions that can be generated locally by two isolated parties using quantum entanglement. (In neither case is there any computational restriction; the only difference is in the kind of shared resource available.) The most famous Bell inequality is the *CHSH inequality*<sup>5</sup>; we will describe another example later, in Section 4.1. The study of Bell inequalities is relevant not only to quantum foundations, where they are a tool to study the nonlocal properties of entanglement, but also in quantum cryptography, where they form the basis for cryptographic protocols, for example, quantum key distribution,<sup>9</sup> and in the study of entangled states of matter in physics. The connection with complexity theory was first made by Cleve<sup>6</sup> using the language of two-player games.

The introduction of entanglement in the setting of interactive proofs has interesting consequences for complexity theory; indeed, it is not *a priori* clear how the class MIP\* compares to MIP. This is because in general the use of entanglement may increase the odds of the provers to convince the verifier of the validity of any given statement, true or false. Such an increase may render a previously sound proof system unsound, because the provers are able to make the verifier accept false statements. Conversely, it is conceivable that new proof systems may be designed for which the completeness property (valid statements have proofs) holds only when the provers are allowed to use entanglement. As a consequence, MIP\* could a priori be smaller, larger, or incomparable to MIP. The only clear inclusions are  $IP \subseteq MIP^*$ , because if the verifier chooses to interact with a single prover, then entanglement does not play any role, and  $MIP^* \subseteq RE$ , the class of recursively enumerable languages, that is languages  $L$  such that there exists a Turing machine  $\mathcal{M}$  such that  $z \in L$  if and only if  $\mathcal{M}$  halts and accepts on input  $z$ . (The inclusion  $MIP^* \subseteq RE$  will be justified once we introduce the model more formally. At the moment we only point out that the

reason that no time-bounded upper bound holds in a self-evident manner is because there is no a priori bound on the complexity of near-optimal provers in an MIP\* interactive proof system. This is in contrast to MIP where any prover can be represented by its question–answer response function, an exponential-size object.)

In<sup>13</sup> the first nontrivial lower bound on MIP\* was obtained, establishing that  $MIP = NEXP \subseteq MIP^*$ . This was shown by arguing that the completeness and soundness properties of the proof system of Babai<sup>1</sup> are maintained in the presence of shared entanglement between the provers. Following<sup>13</sup> a sequence of works established progressively stronger lower bounds, culminating in<sup>18</sup> which showed the inclusion  $NEEXP \subseteq MIP^*$ , where NEEXP stands for nondeterminist doubly exponential time. Since it is known that  $NEXP \subsetneq NEEXP$  it follows that  $MIP \neq MIP^*$ .

Our main result takes this line of work to its limit to show the exact characterization

$$MIP^* = RE.$$

A complete problem for RE is the Halting problem. Thus a surprising consequence of the equality  $MIP^* = RE$  is that there exists a (classical) polynomial-time transformation that takes as input any Turing machine  $\mathcal{M}$  and returns the description of a classical randomized polynomial-time (in the description size of  $\mathcal{M}$ ) verification procedure such that the existence of quantum provers sharing entanglement that are accepted by this verification procedure is equivalent to the fact that the Turing machine halts. Since the actions of quantum provers are (in principle) physically realizable, the verification procedure can be interpreted as the specification of a physical experiment that could be used to certify that a Turing machine halts. The reason that this is surprising is because the Halting problem does not refer to any notion of time bound (and, as shown by Turing, is in fact undecidable), whereas the verification procedure is time-bounded. Yet all true statements (halting Turing machines) have valid proofs in this model, while false statements (non-halting Turing machines) do not.

Before proceeding with a description of the main ideas that go in the proof of this result, we highlight some consequences and describe open questions. Our result is motivated by a connection with Tsirelson’s problem from quantum information theory, itself related to Connes’ Embedding Problem in the theory of von Neumann algebras.<sup>8</sup> In a celebrated sequence of papers, Tsirelson<sup>22</sup> initiated the systematic study of quantum correlation sets, for which he gave two natural definitions. The first definition, referred to as the “tensor product model,” constrains isolated systems to be in tensor product with one another: if two parties Alice and Bob are space-time isolated, then any measurement performed by Alice can be modeled using an operator  $A$  on Alice’s Hilbert space  $\mathcal{H}_A$ , while any measurement performed by Bob can be modeled using an operator  $B$  on Bob’s Hilbert space  $\mathcal{H}_B$ ; studying the correlations between Alice and Bob then involves forming the tensor product  $\mathcal{H}_A \otimes \mathcal{H}_B$  and studying operators such as  $A \otimes B$ . The second definition, referred to as the “commuting model,” is more

general because it only makes use of a single Hilbert space  $\mathcal{H}$  on which both Alice's operators  $A$  and Bob's operators  $B$  act simultaneously; the constraint of "isolation" is enforced by requiring that  $A$  and  $B$  commute,  $AB = BA$ , so that neither operation can have a causal influence on the other.<sup>b</sup> The question of equality between the families of distributions that can be generated in either model is known as *Tsirelson's problem*.<sup>23</sup> As already noted by Fritz,<sup>10</sup> the undecidability of MIP\* implies that Tsirelson's two models are finitely separated (i.e., there is a constant-size family of distributions that can be realized in the second model but is within a constant distance of any distribution that can be realized in the first model); thus, our result that  $\text{RE} \subseteq \text{MIP}^*$  resolves Tsirelson's problem in the negative. Through a sequence of previously known equivalences,<sup>19</sup> we also resolve Connes' Embedding Problem in the negative. As a consequence, our result may lead to the construction of interesting objects in other areas of mathematics. In particular an outstanding open question that may now be within reach is the problem of constructing a finitely presented group that is not sofic, or even not hyperlinear.

### 3. PROOF OVERVIEW

For simplicity, we focus on the class MIP\*(2,1), which corresponds to the setting of one-round protocols with two provers sharing entanglement. (A consequence of our results is that this setting has equal verification power to the general setting of polynomially many provers and rounds of interaction.) The verifier in such a protocol can be described as the combination of two procedures: a *question sampling* procedure  $S$  that samples a pair of questions  $(x, y)$  for the provers according to a distribution  $\mu$  and a *decision* procedure that takes as input the provers' questions and their respective answers  $a, b$  and evaluates a predicate  $D(x, y, a, b) \in \{0, 1\}$  to determine the verifier's acceptance or rejection. (In general, both procedures also take the problem instance  $z$  as input.)

Our results establish the existence of transformations on *families* of two-prover one-round protocols having certain properties. In order to keep track of efficiency (and ultimately, computability) properties, it is important to have a way to specify such families in a uniform manner. Toward this we introduce the following formalism. A *normal form verifier* is specified by a pair of Turing machines  $\mathcal{V} = (\mathcal{S}, \mathcal{D})$  that satisfy certain conditions. The Turing machine  $\mathcal{S}$  (called a *sampler*) takes as input an index  $n \in \mathbb{N}$  and returns the description of a procedure that can be used to sample questions  $(x, y)$  (this procedure itself obeys a certain format associated with "conditionally linear" distributions, defined in Section 4.3 below). The Turing machine  $\mathcal{D}$  (called a *decider*) takes as input an index  $n$ , questions  $(x, y)$ , and answers  $(a, b)$  and returns a single-bit decision. The sampling and decision procedures are required to run in time polynomial in the index  $n$ . Note that normal form verifiers do not take any input other than the index  $n$ ; we explain later in this section how the actual problem instance  $z$  is taken into account. Given a normal form verifier  $\mathcal{V} = (\mathcal{S}, \mathcal{D})$  and an

index  $n \in \mathbb{N}$ , there is a natural two-prover one-round protocol  $\mathcal{V}_n$  associated to it. We let  $\text{val}^*(\mathcal{V}_n)$  denote the maximum success probability of quantum provers sharing entanglement in this protocol.

Our main technical result is a *gap-preserving compression* transformation on normal form verifiers. The following theorem presents an informal summary of the properties of this transformation. For a verifier  $\mathcal{V}_n$  and a probability  $0 \leq p \leq 1$ , we let  $\mathcal{E}(\mathcal{V}_n, p)$  denote the minimum local dimension of an entangled state shared by provers that succeed in the protocol executed by  $\mathcal{V}_n$  with probability at least  $p$ .

**THEOREM 3.1 (GAP-PRESERVING COMPRESSION).** *There exists a polynomial-time Turing machine  $\text{Compress}$  that, when given as input the description of a normal form verifier  $\mathcal{V} = (\mathcal{S}, \mathcal{D})$ , returns the description of another normal form verifier  $\mathcal{V}' = (\mathcal{S}', \mathcal{D}')$  that satisfies the following properties: for all  $n \in \mathbb{N}$ , letting  $N = 2^n$*

1. (**Completeness**) *If  $\text{val}^*(\mathcal{V}_N) = 1$  then  $\text{val}^*(\mathcal{V}'_N) = 1$ .*
2. (**Soundness**) *If  $\text{val}^*(\mathcal{V}_N) \leq \frac{1}{2}$  then  $\text{val}^*(\mathcal{V}'_N) \leq \frac{1}{2}$ .*
3. (**Entanglement**)  *$\mathcal{E}(\mathcal{V}'_N, \frac{1}{2}) \geq \max\{\mathcal{E}(\mathcal{V}_N, \frac{1}{2}), 2^{2^{\Omega(n)}}\}$ .*

The terminology *compression* is motivated by the fact, implicit in the (informal) statement of the theorem, that the time complexity of the verifier's sampling and decision procedures in the protocol  $\mathcal{V}'_n$ , which (as required in the definition of a normal form verifier) is polynomial in  $n$ , is exponentially smaller than the time complexity of the verifier  $\mathcal{V}_N$ , which is polynomial in  $N$  and thus exponential in  $n$ . As such our result can be understood as an efficient delegation procedure for (normal form) interactive proof verifiers. In the next section, we describe how the presence of entanglement between the provers enables such a procedure. First, we sketch how the existence of a Turing machine  $\text{Compress}$  with the properties stated in the theorem implies the inclusion  $\text{RE} \subseteq \text{MIP}^*$ .

To show this, we give an MIP\*(2,1) protocol for the Halting problem, which is a complete problem for RE. Precisely, we give a procedure that given a Turing machine  $\mathcal{M}$  as input returns the description of a normal form verifier  $\mathcal{V}^{\mathcal{M}} = (\mathcal{S}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}})$  with the following properties. First, if  $\mathcal{M}$  does eventually halt on an empty input tape, then it holds that for all  $n \in \mathbb{N}$ ,  $\text{val}^*(\mathcal{V}_n^{\mathcal{M}}) = 1$ . Second, if  $\mathcal{M}$  does not halt then for all  $n \in \mathbb{N}$ ,  $\text{val}^*(\mathcal{V}_n^{\mathcal{M}}) \leq \frac{1}{2}$ . It follows that  $\mathcal{M} \mapsto \mathcal{V}_1^{\mathcal{M}}$  is a valid MIP\*(2,1) proof system for the Halting problem.

We describe the procedure that achieves this. Informally, the procedure returns the specification of a verifier  $\mathcal{V}^{\mathcal{M}} = (\mathcal{S}^{\mathcal{M}}, \mathcal{D}^{\mathcal{M}})$  such that  $\mathcal{D}^{\mathcal{M}}$  proceeds as follows: on input  $(n, x, y, a, b)$  it first executes the Turing machine  $\mathcal{M}$  for  $n$  steps. If  $\mathcal{M}$  halts, then  $\mathcal{D}^{\mathcal{M}}$  accepts. Otherwise,  $\mathcal{D}^{\mathcal{M}}$  computes the description of the compressed verifier  $\mathcal{V}' = (\mathcal{S}', \mathcal{D}')$  that is the output of  $\text{Compress}$  on input  $\mathcal{V}^{\mathcal{M}}$ , then executes the decision procedure  $\mathcal{D}'(n, x, y, a, b)$  and accepts if and only if  $\mathcal{D}'$  accepts.<sup>c</sup>

<sup>b</sup> Precisely, commutation implies that the joint distribution of outcomes obtained by performing one measurement and then the other is independent of the order chosen.

<sup>c</sup> The fact that the decider  $\mathcal{D}^{\mathcal{M}}$  can invoke the  $\text{Compress}$  procedure on itself follows from a well-known result in computability theory known as *Kleene's recursion theorem* (also called *Roger's fixed point theorem*).<sup>15,21</sup>

To show that this procedure achieves the claimed transformation, consider two cases. First, observe that if  $\mathcal{M}$  eventually halts in some number of time steps  $T$ , then by definition  $\text{val}^*(\mathcal{V}_n^{\mathcal{M}}) = 1$  for all  $n \geq T$ . Using Theorem 3.1 along with an inductive argument it follows that  $\text{val}^*(\mathcal{V}_n^{\mathcal{M}}) = 1$  for all  $n \geq 1$ . Second, if  $\mathcal{M}$  never halts, then observe that for any  $n \geq 1$  Theorem 3.1 implies two separate lower bounds on the amount of entanglement required to win the protocol  $\mathcal{V}_n^{\mathcal{M}}$  with probability at least  $\frac{1}{2}$ : the dimension is (a) at least  $2^{2^{2^{(n)}}}$  and (b) at least the dimension needed to succeed in the protocol  $\mathcal{V}_{2^n}^{\mathcal{M}}$  with probability at least  $\frac{1}{2}$ . By induction, it follows that an *infinite* amount of entanglement is needed to succeed in the protocol  $\mathcal{V}_n$  with any probability greater than  $\frac{1}{2}$ . By continuity, a sequence of finite-dimension prover strategies for  $\mathcal{V}_n$  cannot lead to a limiting value larger than  $\frac{1}{2}$ , and  $\text{val}^*(\mathcal{V}_n^{\mathcal{M}}) \leq \frac{1}{2}$ .

#### 4. USING ENTANGLEMENT TO COMPRESS QUANTUM INTERACTIVE PROOFS

In the language introduced in the previous section, the inclusion  $\text{NEXP} \subseteq \text{MIP}$  implies that for any  $\text{NEXP}$ -complete language  $L$ , there is a pair of polynomial-time Turing machines  $\mathcal{V} = (\mathcal{S}, \mathcal{D})$  such that the following hold. The Turing machine  $\mathcal{S}$  takes as input an integer  $n \in \mathbb{N}$  and returns a (probabilistic) circuit that can be used to sample questions  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  according to some distribution  $\mathcal{S}_n$ . The Turing machine  $\mathcal{D}$  on input  $n \in \mathbb{N}$  as well as an instance  $z \in \{0, 1\}^n$  and  $(x, y, a, b) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \mathcal{B}$  returns a decision bit in  $\{0, 1\}$ . This proof system is such that whenever  $z \in L$  there are “proofs”  $f_A : \mathcal{X} \rightarrow \mathcal{A}$  and  $f_B : \mathcal{Y} \rightarrow \mathcal{B}$  and such that for all  $(x, y)$  in the support of  $\mathcal{S}_n$ ,  $\mathcal{D}(1^n, z, x, y, f_A(x), f_B(y)) = 1$ , and whenever  $z \notin L$  then for any purported “proofs”  $f_A : \mathcal{X} \rightarrow \mathcal{A}$  and  $f_B : \mathcal{Y} \rightarrow \mathcal{B}$  it holds that

$$\mathbb{E}_{(x,y) \sim \mathcal{S}_n} \mathcal{D}(1^n, z, x, y, f_A(x), f_B(y)) \leq \frac{1}{2}. \quad (1)$$

While the definition of  $\text{MIP}$  allows for more general proof systems, for example, the sampling of questions may depend on the input  $z$  and there may be multiple rounds of interaction, subsequent refinements of the original proof that  $\text{NEXP} \subseteq \text{MIP}$  imply that proof systems of the form above are sufficient to capture the entire class.

The maximum of the expression on the left-hand side of (1) can be computed exactly in nondeterministic exponential time by guessing an optimal choice of  $(f_A, f_B)$  and evaluating the expression. Therefore, the class  $\text{MIP}$  does not allow verification of languages beyond  $\text{NEXP}$ . In particular, Theorem 3.1 is clearly not possible in this model, because by applying it to the verifier for  $\text{NEXP}$  described in the preceding paragraph one would obtain a polynomial-time verifier with the ability to decide a complete language for  $\text{NEXP}$ ; however, it is known that  $\text{NEXP} \not\subseteq \text{NEXP}$ . To go beyond  $\text{NEXP}$  and prove Theorem 3.1, we must find ways for the verifier to constrain the provers to make use of entanglement so that even more complex computations can be delegated to them.

To prove Theorem 3.1, we will show how the actions of both the sampler  $\mathcal{S}$  and the decider  $\mathcal{D}$  can be delegated to the provers and their correct execution verified in time polylogarithmic in the sampler and decider’s time complexity. This will enable a polynomial-time verifier to force the

provers to “simulate” the action of  $\mathcal{S}$  and  $\mathcal{D}$  on inputs of size  $N = 2^n$  using polynomial resources, as required by Theorem 3.1. While the techniques for delegating deterministic procedures such as  $\mathcal{D}$  are already present, to some extent, in the proof of  $\text{MIP} = \text{NEXP}$  (whose implications for delegated computation are well-known, see, e.g., Goldwasser<sup>12</sup>), for the inherently probabilistic procedure  $\mathcal{S}$ , we are faced with an entirely new challenge: the delegation of *randomness generation*. In particular, it is clearly not reasonable to give the provers access to the random seed used by  $\mathcal{S}$ , as this would provide them perfect knowledge of both questions and therefore allow them to easily defeat the protocol. In order to ensure that the provers sample from the correct distribution while each obtaining exactly the right amount of information about their respective input, we leverage entanglement in an essential way. In the next section, we give a classic example of how randomness can be certified by using entanglement. In Section 4.2, we introduce an asymptotically efficient entanglement test that builds on the example. In Section 4.3, we describe a class of question distribution that can be delegated (we will say “introspected”) by making use of the entanglement test. In Section 4.5, we explain a final step of parallel repetition, and in Section 4.6, we conclude.

#### 4.1. The Magic Square game

We introduce a two-player game known as the “Magic Square game.” To specify the rules of the game, we first describe a (classical) constraint satisfaction problem that consists of 6 linear equations defined over 9 variables that take values in the binary field  $\mathbb{F}_2$ . The variables are associated with the cells of a  $3 \times 3$  grid, as depicted in Figure 1. Five of the equations correspond to the constraint that the sum of the variables in each row and the first two columns must be equal to 0, and the last equation imposes that the sum of the variables in the last column must be equal to 1.

This system of equations is clearly unsatisfiable. Now consider the following game associated to it. In the game, the trusted referee first samples one of the 6 equations uniformly at random and then one of the three variables appearing in the constraint uniformly at random. It asks one prover for an assignment of values to all variables in the chosen constraint, and the other prover for the selected variable. It is not hard to see that no deterministic or randomized strategy can succeed in this game with probability larger than  $\frac{17}{18}$ , since there are 18 pairs of questions in total and any deterministic or randomized strategy that succeeds with a strictly larger probability is easily seen to imply a satisfying assignment to the Magic Square.

What makes the game “magic” is the surprising fact, first demonstrated by Mermin and Peres,<sup>16, 20</sup> that this game has

Figure 1. The Magic Square game.

$x_1$	$x_2$	$x_3$
$x_4$	$x_5$	$x_6$
$x_7$	$x_8$	$x_9$



a perfect quantum strategy: two noncommunicating players sharing entanglement can win with probability 1. Moreover, this can be achieved by sharing a simple 4-dimensional quantum state (two EPR pairs) and making local measurements on it (i.e., each prover only makes an observation on their local share of the state; see Section 4.2 for the mathematical formalism of quantum strategies). The outcomes of local measurements are not causally related and entanglement does not allow the provers to communicate information. However, their outcome distribution can be correlated: in the Mermin-Peres quantum strategy for the Magic Square game, for any pair of questions selected by the referee the joint distribution of the player’s answers happens to be uniform over all pairs of possible answers.<sup>d</sup>

The Magic Square game is an example of a Bell inequality (specifically, the inequality is that classical strategies cannot win with probability larger than  $\frac{17}{18}$ ). The game can be interpreted as a two-prover verification protocol for the satisfiability of the underlying system of 9 equations, in which case the protocol is sound with classical provers but unsound with quantum provers. However, for our purposes, this is not a productive observation—we are interested in finding *positive* uses for entanglement. Is there anything useful that can be certified using this game?

We reproduce a crucial observation due to<sup>7</sup>: the *only* way that quantum provers are able to succeed in the Magic Square game beyond what is possible classically is by generating answers that are *intrinsically random*. This is because an intrinsically random strategy (by which we mean that the computation of each answer *must* involve injecting fresh randomness, that was not determined prior to the questions being generated) is the only possibility for evading the classical impossibility (the aforementioned “Bell inequality”). Indeed if the two players always give a deterministic and valid answer to their questions then listing those 18 pairs of answers will lead to a satisfying assignment for the magic square constraints. Since this is not possible it must be that each pair of answers is generated “on the fly,” so that any two answers are correlated in the right way but there is no meaningful correlation between pairs of answers obtained to different pairs of questions (e.g., in different runs of the game). In other words, quantum randomness cannot be “fixed,” and this is what allows quantum provers to succeed in the game.

Based on this observation and with further technical work, it is possible to use the Magic Square game to develop a simple two-prover test (i.e., a small self-contained protocol that can be used as a building block towards a larger MIP\* protocol) that constrains the provers to obtain, and report to the verifier, identical uniformly random outcomes. Note that this is stronger than merely allowing the provers to use shared randomness, in the sense that it is *verifiable*: upon seeing either prover’s answer, the verifier has the guarantee that it has been chosen near-uniformly at random—the provers cannot bias the randomness in any way (unless they pay a price by failing in the test with positive probability).

<sup>d</sup> Not every game has such a perfect quantum strategy. For example, a game in which the second player has to answer the first player’s question: this would violate the nonsignaling principle.

While this is already a genuinely quantum possibility that transforms the realm of achievable protocols, it does not suffice for our purposes, for two reasons. First of all the distributions that we aim to sample from, that is the distributions  $S_n$ , are more complicated than uniform shared randomness. We discuss this point in detail in Section 4.3. Second, executing as many copies of the Magic Square game as there are bits of randomness required to sample from  $S_n$  would not lead to the complexity savings that we are aiming to achieve. This problem can be solved by employing a more efficient means of randomness and entanglement generation that builds simultaneously on the Magic Square game and on a quantum version of the classical low-degree test, a key component in the proof of  $\text{NEXP} \subseteq \text{MIP}$ , which we repurpose for this new goal. We explain this in the next section.

## 4.2. The quantum low-degree test

The quantum low-degree test, first introduced by Natarajan<sup>17</sup> and analyzed by Ji,<sup>14</sup> is one of the core technical components behind the proof of Theorem 3.1. The test provides an efficient means of certifying entanglement (and, as a corollary, randomness generation) between two provers. The test builds upon classical results in the property testing and probabilistically checkable proofs literature for testing low-degree polynomials, a line of work initiated by Gemmel<sup>11</sup> building upon the multilinearity test by Babai.<sup>1</sup> In this section, we first state the “quantum” version of these results and then explain its use for delegating the sampling procedure.

To state the quantum low-degree test in sufficient technical depth, we introduce the standard formalism used to model quantum strategies. A strategy for two quantum provers (in a one-round protocol) consists of (i) a quantum state shared by the provers, which is represented by a unit vector  $|\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B$  where  $\mathcal{H}_A$  and  $\mathcal{H}_B$  are finite-dimensional Hilbert spaces associated with each prover (e.g., if  $|\psi\rangle$  consists of  $n$  EPR pairs then  $\mathcal{H}_A = \mathcal{H}_B = \mathbb{C}^{2^n}$ ) and (ii) for each question to a prover, say  $x \in \mathcal{X}$  to the first prover, a measurement  $\{A_a^x\}_{a \in \mathcal{A}}$  that the prover performs on their share of the quantum state to produce an answer  $a \in \mathcal{A}$ . Mathematically each  $A_a^x$  is a positive semidefinite operator acting on  $\mathcal{H}_A$  such that  $\sum_{a \in \mathcal{A}} A_a^x = I$ . The measurement rule states that two provers modeled in this way generate answers  $(a, b)$  in response to questions  $(x, y)$  with probability

$$p(a, b | x, y) = \langle \psi | A_a^x \otimes B_b^y | \psi \rangle.$$

Clearly two strategies generate the same distributions if they are unitarily equivalent, that is, exchanging  $|\psi\rangle$  for  $U_A \otimes U_B |\psi\rangle$  for unitaries  $U_A$  on  $\mathcal{H}_A$  and  $U_B$  on  $\mathcal{H}_B$  and conjugating all operators by  $U_A$  or  $U_B$  leads to the same strategy. So any characterization of near-optimal strategies will only be “up to local unitaries.” More generally, the provers may have access to a larger space that is not directly used in their strategy. For this reason, a characterization of near-optimal quantum strategies is generally formulated “up to local isometries,” where a local isometry is a linear map  $\phi_A : \mathcal{H}_A \rightarrow \mathcal{H}_{A'} \otimes \mathcal{H}_{A''}$  (resp.  $\phi_B : \mathcal{H}_B \rightarrow \mathcal{H}_{B'} \otimes \mathcal{H}_{B''}$ ).

The quantum low-degree test comes in the form of a *rigidity* result, which guarantees that any near-optimal strategy

in a certain two-prover protocol is essentially unique—up to local isometries. The test is parametrized by a tuple  $\text{qldparams} = (q, m, d)$  where  $q$  is a field size,  $m$  a number of variables, and  $d$  a degree satisfying some conditions that we omit from this abstract. We denote the associated two-prover one-round verifier  $\mathcal{V}_{\text{qldparams}}^{\text{PAULI}}$ . In the associated protocol, whose exact description we also omit, the provers are expected to (i) measure  $2^m \log q$  shared EPR pairs in a certain basis (either the computational basis or the Hadamard basis, which is its Fourier transform), obtaining an outcome  $a \in \mathbb{F}_q^{2^m}$ ; (ii) interpolate a polynomial  $g_a : \mathbb{F}_q^m \rightarrow \mathbb{F}_q$  of individual degree at most  $d$ ; and (iii) return the restriction of  $g_a$  to a line or point in  $\mathbb{F}_q^m$  that is provided to them as their question. This is almost exactly the same as the classic test from Babai<sup>1</sup> and Gemmel<sup>11</sup>; the “quantum” part of the test lies in the fact that the specification  $a$  for the polynomial  $g_a$  used by the provers should be obtained as a result of a measurement on an entangled state. This measurement can be required to be made in different, incompatible bases (depending on the question), and thus its outcome cannot be fixed a priori, before the protocol starts. (In contrast, in the classical case, it is assumed that  $a$  is a fixed string on which the provers pre-agree.) Tests based on the Magic Square game are performed in addition to the above to enforce that the provers make their measurements in the right bases.

**THEOREM 4.1.** *There exists a function*

$$\delta_{\text{QLD}}(\varepsilon, q, m, d) = a(md)^a (\varepsilon^b + q^{-b} + 2^{-bmd})$$

for universal constants  $a \geq 1$  and  $0 < b < 1$  such that the following holds. For all admissible parameter tuples  $\text{qldparams} = (q, m, d)$  and for all strategies  $\mathcal{S} = (|\psi\rangle, A, B)$  for  $\mathcal{V}_{\text{qldparams}}^{\text{PAULI}}$  that succeed with probability at least  $1 - \varepsilon$ , there exist local isometries  $\phi_A : \mathcal{H}_A \rightarrow \mathcal{H}_{A'} \otimes \mathcal{H}_{A''}$ ,  $\phi_B : \mathcal{H}_B \rightarrow \mathcal{H}_{B'} \otimes \mathcal{H}_{B''}$  and a state  $|\text{AUX}\rangle \in \mathcal{H}_{A'} \otimes \mathcal{H}_{B'}$  such that

$$\| |\phi_A \otimes \phi_B |\psi\rangle - |\text{AUX}\rangle \otimes |\text{EPR}\rangle^{\otimes M \log q} \| \leq \delta_{\text{QLD}}(\varepsilon, q, m, d),$$

where  $|\text{EPR}\rangle$  denotes an EPR pair. Moreover, letting  $\tilde{A}_a^x = \phi_A^x A_a^x \phi_A^\dagger$  and  $\tilde{B}_a^y = \phi_B^y B_a^y \phi_B^\dagger$  we have for  $W \in \{X, Z\}$

$$\begin{aligned} \tilde{A}_u^W \otimes I_{B'B''} &\approx_{\delta_{\text{QLD}}} (\sigma_u^W)_{A'} \otimes I_{A''B''} \\ I_{A'A''} \otimes \tilde{B}_u^W &\approx_{\delta_{\text{QLD}}} I_{A'A''B'} \otimes (\sigma_u^W)_{B''}. \end{aligned}$$

In the theorem statement, the approximation  $\approx_{\delta_{\text{QLD}}}$  is measured in a norm that depends on the state shared by the provers and is appropriate for arguing that the two measurement operators on the left and right of the approximation sign lead to similar outcome distributions, even when some other operator is applied by the other prover. The measurement operators  $A_u^W$  and  $B_u^W$  on the left-hand side are the provers’ measurement operators associated with a designated pair of questions  $W = X, Z$  in the protocol. The operators  $\sigma_u^W$  that appear on the right-hand side denote tensor products of single-qubit Pauli measurements in the basis  $W$ , which is the computational basis in case  $W = Z$  and the Hadamard basis in case  $W = X$ . Here  $u$  is an element of  $\mathbb{F}^{2^m}$ , which can be interpreted as an  $(2^m \log q)$ -bit string.

Informally, the theorem guarantees that any prover strategy that achieves a high probability of success in the test is “locally equivalent” to a strategy that consists in measuring  $2^m \log q$  EPR pairs in the appropriate basis, which is a strategy of the “honest” form described above. Crucially, the number of EPR pairs tested is exponential in the verifier complexity, which scales polynomially in  $\log q, m$ , and  $d$ .

We turn our attention back to the problem of delegating the sampling from the distribution  $\mathcal{S}_n$  to the provers. Referring to the provers producing their own questions, we call this procedure “introspection.” In general, it is not known how to introspect arbitrary unstructured distributions  $\mathcal{S}_n$ . The verifier that underlies the protocol from Babai<sup>1</sup> chooses most of its question from the “line-point” distribution  $\mathcal{S}_n$ , which is the distribution over pairs  $(x, y)$  such that  $x$  is the description of a uniformly random line  $\ell \in \mathbb{F}_q^m$  and  $y$  is a uniformly random point on  $\ell$ . It is natural to first consider introspection of this distribution. This is done by Natarajan<sup>18</sup> to “compress” the protocol from Babai<sup>1</sup> and thereby show the inclusion  $\text{NEEXP} \subseteq \text{MIP}^*$ . The compressed protocol relies on the test from Theorem 4.1 to efficiently sample its questions to the provers; the randomness used by the provers for sampling is based on their measurement outcomes in the test, which are bitstrings of length  $2^m \log q$ .

Unfortunately the distribution used to sample the questions in the compressed protocol no longer has such a nice form as the lines-point distribution. In particular, it includes the question distribution from the quantum low-degree test, which itself incorporates the Magic Square game question distribution. Since our goal is to compress protocols iteratively (recall the end of Section 3), it is essential to identify a class of distributions that is sufficiently general and “closed under introspection” in the sense that any distribution from the family can be tested using a distribution from the same family with reduced randomness complexity. For this, we introduce the class of *conditionally linear distributions*, which generalizes the low-degree test distribution. We show that conditionally linear distributions can be “introspected” using conditionally linear distributions only, enabling recursive introspection. (As we will see later, other closure properties of conditionally linear distributions, such as taking direct products, play an important role as well.) We describe this family of distributions next.

### 4.3. Conditionally linear distributions

Fix a vector space  $V$  that is identified with  $\mathbb{F}_q^m$ , for a finite field  $\mathbb{F}_q$  and integer  $m$ . Informally, a function  $L$  on  $V$  is *conditionally linear* (CL for short) if it can be evaluated by a procedure that takes the following form: (i) read a substring  $z^{(1)}$  of  $z$ ; (ii) evaluate a linear function  $L_1$  on  $z^{(1)}$ ; and (iii) repeat steps (i) and (ii) with the remaining coordinates  $z \setminus z^{(1)}$ , such that the next steps are allowed to depend in an arbitrary way on  $L_1(z^{(1)})$  but not directly on  $z^{(1)}$  itself. What distinguishes a function of this form from an arbitrary function is that we restrict the number of iterations of (i)—(ii) to a constant number (at most 9, in our case). (One may also think of CL functions as “adaptively linear” functions, where the number of “levels” of adaptivity is the number of iterations of (i)—(ii).) A distribution  $\mu$  over pairs  $(x, y) \in V \times V$  is called conditionally linear

if it is the image under a pair of conditionally linear functions  $L^A, L^B : V \rightarrow V$  of the uniform distribution on  $V$ , that is,  $(x, y) \sim (L^A(z), L^B(z))$  for uniformly random  $z \in V$ .

An important class of CL distributions are low-degree test distributions, which are distributions over question pairs  $(x, y)$  where  $y$  is a randomly chosen affine subspace of  $\mathbb{F}_q^m$  and  $x$  is a uniformly random point on  $y$ . We explain this for the case where the random subspace  $y$  is one-dimensional (i.e., a line). Let  $V = V_x \otimes V_v$  where  $V_x = V_v = \mathbb{F}_q^m$ . Let  $L^A$  be the projection onto  $V_x$  (i.e., it maps  $(x, v) \rightarrow (x, 0)$  where  $x \in V_x$  and  $v \in V_v$ ). Define  $L^B : V \rightarrow V$  as the map  $(x, v) \mapsto (L_v^{L^A}(x), v)$  where  $L_v^{L^A} : V_x \rightarrow V_x$  is a linear map that, for every  $v \in V_v$ , projects onto a complementary subspace to the one-dimensional subspace of  $V_x$  spanned by  $v$  (one can think of this as an “orthogonal subspace” to the span of  $\{v\}$ ).  $L^B$  is conditionally linear because it can be seen as first reading the substring  $v \in V_v$  (which can be interpreted as specifying the *direction* of a line), and then applying a linear map  $L_v^{L^A}$  to  $x \in V_x$  (which can be interpreted as specifying a canonical point on the line  $\ell = \{x + tv : t \in \mathbb{F}\}$ ). It is not hard to see that the distribution of  $(L^A(z), L^B(z))$  for  $z$  uniform in  $V$  is identical (up to relabeling) to the low-degree test distribution  $(x, \ell)$  where  $\ell$  is a uniformly random affine line in  $\mathbb{F}_q^m$ , and  $x$  is a uniformly random point on  $\ell$ .

We show that any CL distribution  $\mu$ , associated with a pair of CL functions  $(L^A, L^B)$  over a linear space  $V = \mathbb{F}_q^m$ , can be “introspected” using a CL distribution that is “exponentially smaller” than the initial distribution. Slightly more formally, to any CL distribution  $\mu$ , we associate a two-prover test in which questions from the verifier are sampled from a CL distribution  $\mu'$  over  $\mathbb{F}_q^{m'}$  for some  $m' = \text{poly} \log(m)$  and such that in any successful strategy, when the provers are queried on a special question labeled **INTRO** they must respond with a pair  $(x, y)$  that is approximately distributed according to  $\mu$ . (The test allows us to do more: it allows us to conclude how the provers obtained  $(x, y)$ —by measuring shared EPR pairs in a specific basis—and this will be important when using the test as part of a larger protocol that involves other checks.) Crucially for us, the distribution  $\mu'$  only depends on a size parameter associated with  $(L^A, L^B)$  (essentially, the integer  $m$  together with the number of “levels” of adaptivity of  $L^A$  and  $L^B$ ), but not on any other structural property of  $(L^A, L^B)$ . Only the decision predicate for the associated protocol depends on the entire description of  $(L^A, L^B)$ .

We say a few words about the design of  $\mu'$  and the associated test, which borrow heavily from Natarajan.<sup>18</sup> Building on the quantum low-degree test introduced in Section 4.2, we already know how a verifier can force a pair of provers to measure  $m$  EPR pairs in either the computational or Hadamard basis and report the (necessarily identical) outcome  $z$  obtained, all the while using questions of length polylogarithmic in  $m$  only. The added difficulty is to ensure that a prover obtains, and returns, precisely the information about  $z$  that is contained in  $L^A(z)$  (resp.  $L^B(z)$ ), and not more. A simple example is the line-point distribution described earlier: there, the idea to ensure that, for example, the “point” prover only obtains the first component,  $x$  of  $(x, v) \in V_x \otimes V_v$ , the verifier demands that the “point” prover measures their qubits associated with the space  $V_v$

in the Hadamard, instead of computational, basis; due to the uncertainty principle, this has the effect of “erasing” the outcome in the computational basis. The case of the “line” prover is a little more complex: the goal is to ensure that, conditioned on the specification of the line  $\ell$  received by the “line” prover, the point  $x$  received by the “point” prover is uniformly random within  $\ell$ . This was shown to be possible in [18].

#### 4.4. Answer reduction

The previous section sketches how we are able to use entanglement testing techniques to delegate the sampling of questions directly to the provers, with an exponential savings in the verifier’s effort. Let  $\mathcal{V}_n^{\text{INTRO}}$  denote the resulting “question-reduced” verifier. However, the players still respond with poly( $N$ )-length answers, which the verifier has to check satisfies the decision predicate of the original protocol. We explain how to obtain a similar exponential savings in the verification of the answers.

For this, we use probabilistically checkable proofs (PCPs) as follows. The verifier in  $\mathcal{V}_n^{\text{AR}}$  samples questions as  $\mathcal{V}_n^{\text{INTRO}}$  would and sends them to the players. Instead of receiving the introspected questions and answers  $(x, y, a, b)$  for the original verifier  $\mathcal{V}_n$  and running the decision procedure  $\mathcal{D}(N, x, y, a, b)$ , the verifier instead asks the provers to compute a PCP  $\Pi$  for the statement that the original decider  $\mathcal{D}$  accepts the input  $(N, x, y, a, b)$  in time  $T = \text{poly}(N)$ . The verifier then samples additional questions for the provers that ask them to return specific entries of the proof  $\Pi$ . Finally, upon receipt of the provers’ answers, the verifier executes the PCP verification procedure. Because of the efficiency of the PCP, both the sampling of the additional questions and the decision procedure can be executed in time poly( $n$ ).<sup>c</sup>

This sketch presents some difficulties. A first difficulty is that in general no prover by themselves has access to the entire input  $(N, x, y, a, b)$  to  $\mathcal{D}$ , so no prover can compute the entire proof  $\Pi$ . This issue is addressed using oracularization, which is a classical technique and so we omit the details (there are some subtleties specific to the quantum case). A second difficulty is that a black-box application of an existing PCP, as done by Natarajan,<sup>18</sup> results in a question distribution for  $\mathcal{V}_n^{\text{AR}}$  (i.e., the sampling of the proof locations to be queried) that is rather complex—and in particular, it may no longer fall within the framework of CL distributions for which we can do introspection. To avoid this, we design a bespoke PCP based on the classical MIP for NEXP (in particular, we borrow and adapt techniques from Ben-Sasson<sup>3,4</sup>). Two essential properties for us are that (i) the PCP proof is a collection of several low-degree polynomials, two of which are low-degree encodings of each prover’s uncompressed answers, and (ii) verifying the proof only requires (a) running low-degree tests, (b) querying all polynomials at a uniformly random point, and (c) performing simple consistency checks. Property (i) allows us to eliminate the extra layer of encoding by Natarajan,<sup>18</sup> who had to consider a PCP

<sup>c</sup> This idea is inspired by the technique of composition in the PCP literature, in which the complexity of a verification procedure can be reduced by composing a proof system (often a PCP itself) with another PCP.

of proximity for a circuit applied to the low-degree encodings of the provers' uncompressed answers. Property (ii) allows us to ensure the question distribution employed by the verifier remains conditionally linear.

#### 4.5. Parallel repetition

The combined steps of question reduction (via introspection) and answer reduction (via PCP composition) result in a protocol whose verifier  $\mathcal{V}_n^{\text{AR}}$  has complexity poly( $n$ ). Furthermore, the property of having a perfect quantum strategy is preserved by the reduction. Unfortunately, the sequence of transformations incurs a loss in the soundness parameters: if  $\text{val}^*(\mathcal{V}_n) \leq \frac{1}{2}$ , then we can only establish that  $\text{val}^*(\mathcal{V}_n^{\text{AR}}) \leq 1 - C$  for some positive constant  $C < \frac{1}{2}$  (we call  $C$  the soundness gap). Such a loss would prevent us from recursively applying the compression procedure Compress an arbitrary number of times, which is needed to obtain the desired complexity results for MIP\*.

To overcome this, we need a final transformation to restore the soundness gap after answer reduction to a constant larger than  $\frac{1}{2}$ . To achieve this, we use the technique of parallel repetition. The parallel repetition of a two-prover one-round protocol  $\mathcal{G}$  is another protocol  $\mathcal{G}^k$ , for some number of repetitions  $k$ , which consists of executing  $k$  independent and simultaneous instances of  $\mathcal{G}$  and accepting if and only if all  $k$  instances accept. Intuitively, parallel repetition is meant to decrease the prover's maximum success probability in a protocol  $\mathcal{G}$  exponentially fast in  $k$ , provided  $\text{val}^*(\mathcal{G}) < 1$  to begin with. However, it is an open question of whether this is generally true for the entangled value  $\text{val}^*$ .

Nevertheless, some variants of parallel repetition are known to achieve exponential amplification. We use a variant called "anchored parallel repetition" introduced by Bavarian.<sup>2</sup> This allows us to devise a transformation that efficiently amplifies the soundness gap to a constant. The resulting protocol  $\mathcal{V}_n^{\text{REP}}$  has the property that if  $\text{val}^*(\mathcal{V}_n^{\text{AR}}) = 1$ , then  $\text{val}^*(\mathcal{V}_n^{\text{REP}}) = 1$  (and moreover this is achieved using a commuting and consistent strategy), whereas if  $\text{val}^*(\mathcal{V}_n^{\text{AR}}) \leq 1 - C$  for some universal constant  $C > 0$ , then  $\text{val}^*(\mathcal{V}_n^{\text{REP}}) \leq \frac{1}{2}$ . Furthermore, we have the additional property, essential for us, that good strategies in  $\mathcal{V}_n^{\text{REP}}$  require as much entanglement as good strategies in  $\mathcal{V}_n^{\text{AR}}$  (which in turn require as much entanglement as good strategies in  $\mathcal{V}_n$ ). The complexity of the verifier in  $\mathcal{V}_n^{\text{REP}}$  remains poly( $n$ ).

The anchored parallel repetition procedure, when applied to a normal form verifier, also yields a normal form verifier: this is because the direct product of CL distributions is still conditionally linear.

#### 4.6. Putting it all together

This completes the overview of the transformations performed by the compression procedure Compress of Theorem 3.1. To summarize, given an input normal form verifier  $\mathcal{V}$ , question and answer reduction are applied to obtain  $\mathcal{V}^{\text{AR}}$ , and anchored parallel repetition is applied to obtain  $\mathcal{V}^{\text{REP}}$ , which is returned by the compression procedure. Each of these transformations preserves completeness (including the commuting and consistent properties of a value-1 strategy) as well as the entanglement requirements of each protocol; moreover, the overall transformation preserves soundness. □

#### References

- Babai, L., Fortnow, L., Lund, C. Non-deterministic exponential time has two-prover interactive protocols. *Comput. Complex. 1*, 1 (1991), 3–40.
- Bavarian, M., Vidick, T., Yuen, H. Hardness amplification for entangled games via anchoring. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ACM, New York, NY, USA, 2017, 303–316.
- Ben-Sasson, E., Goldreich, O., Harsha, P., Sudan, M., Vadhan, S. Robust PCPs of proximity, shorter PCPs, and applications to coding. *SIAM J. Comput. 36*, 4 (2006), 889–974.
- Ben-Sasson, E., Sudan, M. Simple PCPs with poly-log rate and query complexity. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, ACM, New York, NY, USA, 2005, 266–275.
- Clouser, J.F., Horne, M.A., Shimony, A., Holt, R.A. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett. 23*, 15 (1969), 880.
- Cleve, R., Hoyer, P., Toner, B., Watrous, J. Consequences and limits of nonlocal strategies. In *Proceedings. 19th IEEE Annual Conference on Computational Complexity, 2004*, IEEE, Washington, DC, USA, 2004, 236–249.
- Colbeck, R. Quantum and relativistic protocols for secure multi-party computation. Ph. D. Thesis, 2009.
- Connes, A. Classification of injective factors cases II<sub>1</sub>, II<sub>∞</sub>, III<sub>λ</sub>, λ ≠ 1. *Ann. Math. 104* (1976), 73–115.
- Ekert, A.K. Quantum cryptography based on bell's theorem. *Phys. Rev. Lett. 67*, 6 (1991), 661.
- Fritz, T., Netzer, T., Thom, A. Can you compute the operator norm? *Proc. Am. Math. Soc. 142*, 12 (2014), 4265–4276.
- Gemmel, P., Lipton, R., Rubinfeld, R., Sudan, M., Wigderson, A. Self testing/correcting for polynomials and for approximate functions. In *Proceedings of the 23rd STOC*. ACM, New York, NY, USA, 1991, 32–42.
- Goldwasser, S., Kalai, Y.T., Rothblum, G.N. Delegating computation: interactive proofs for muggles. *J. ACM 62*, 4 (2015), 1–64.
- Ito, T., Vidick, T. A multi-prover interactive proof for NEXP sound against entangled provers. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, IEEE, Washington, DC, USA, 2012, 243–252.
- Ji, Z., Natarajan, A., Vidick, T., Wright, J., Yuen, H. Quantum soundness of the classical low individual degree test. arXiv preprint arXiv:2009.12982 (2020).
- Kleene, S.C. Introduction to Metamathematics. *J. Symb. Log. 19*, 3 (1954), 215–216.
- Mermin, D. Simple unified form for the major no-hidden-variables theorems. *Phys. Rev. Lett. 65*, 27 (1990), 3373.
- Natarajan, A., Vidick, T. Two-player entangled games are NP-hard. In *Proceedings of the 33rd Computational Complexity Conference*, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2018, 20.
- Natarajan, A., Wright, J. NEXP ⊆ MIP\*. arXiv preprint arXiv:1904.05870v3 (2019).
- Ozawa, N. About the Connes embedding conjecture. *Jpn. J. Math. 8*, 1 (2013), 147–183.
- Peres, A. Incompatible results of quantum measurements. *Phys. Lett. A 151*, 3–4 (1990), 107–108.
- Rogers Jr., H. *Theory of Recursive Functions and Effective Computability*. MIT Press, Cambridge, MA, USA, 1987.
- Tsirelson, B.S. Some results and problems on quantum bell-type inequalities. *Hadronic J. Suppl. 8*, 4 (1993), 329–345.
- Tsirelson, B.S. Bell inequalities and operator algebras, 2006. Problem statement from website of open problems at TU Braunschweig, 2006. Available at <http://web.archive.org/web/20090414083019/http://www.imaph.tu-bs.de/qi/problems/33.html>.

**Zhengfeng Ji** (zhengfeng.ji@uts.edu.au), School of Computer Science, University of Technology Sydney, Sydney, Australia.

**Anand Natarajan** (anandn@caltech.edu), Department of EECS, Massachusetts Institute of Technology, Cambridge, MA, USA.

**Thomas Vidick** (vidick@caltech.edu), Department of Computing and

Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA.

**John Wright** (wright@cs.utexas.edu), Department of Computer Science, University of Texas at Austin, Austin, TX, USA.

**Henry Yuen** (hyuen@cs.columbia.edu), Department of Computer Science, Columbia University, New York, NY, USA.

## **Johns Hopkins University** *Lecturer/Sr. Lecturer in Computer Science*

The Department of Computer Science at Johns Hopkins University seeks applicants for full-time teaching positions. These are career-oriented, renewable appointments, responsible for the development and delivery of undergraduate and/or graduate courses, depending on the candidate's background. We are searching broadly to meet teaching needs across the discipline, including data science and machine learning. Each position carries a 3-course load per semester, usually with only 2 different preps. Teaching faculty are encouraged to engage in educational research and departmental and university service and may have advising responsibilities. Extensive grading support is given to all instructors. The university has instituted a non-tenure track career path for full-time teaching faculty culminating in the rank of Teaching Professor.

Johns Hopkins is a private university known for its commitment to academic excellence and research. The Computer Science department is one of nine academic departments in the Whiting School of Engineering, on the beautiful Homewood Campus. We are located in Baltimore, MD in close proximity to Washington, DC and Philadelphia, PA. See the department webpage at <https://cs.jhu.edu> for additional information about the department, including undergraduate and graduate programs and current course descriptions.

Applications may be submitted online at <http://apply.interfolio.com/94564>. Questions may be directed to [lecsearch2021@cs.jhu.edu](mailto:lecsearch2021@cs.jhu.edu). For full consideration, applications should be submitted by December 1, 2021. Applications will be accepted until the position is filled.

The Department is conducting a broad and inclusive search and is committed to identifying candidates who through their teaching and service will contribute to the diversity and excellence of the academic community.

The Johns Hopkins University is committed to active recruitment of a diverse faculty and student body. The University is an Affirmative Action/Equal Opportunity Employer of women, minorities, protected veterans and individuals with disabilities and encourages applications from these and other protected group members. Consistent with the University's goals of achieving excellence in all areas, we will assess the comprehensive qualifications of each applicant.

### **Requirements**

Applicants for the position should have a Ph.D. in Computer Science or a closely related field. Demonstrated excellence in and commitment to teaching, and excellent communication skills are expected of all applicants.

## **San Diego State University** **Department of Computer Science** *Tenure-Track Assistant Professor*

Tenure-Track Assistant Professor Position The Department of Computer Science is seeking to hire a tenure-track assistant professor beginning Fall 2022. Strong candidates in all fields of computer science will be considered, with an emphasis on software engineering. The candidates should have a PhD degree in Computer Science or a closely related field. Position details and instructions to apply can be found at <https://apply.interfolio.com/94563>. Questions about the position may be directed to [COS-CS-SE-Search2022@sdsu.edu](mailto:COS-CS-SE-Search2022@sdsu.edu). SDSU is an equal opportunity/Title IX employer.

## **University of Central Missouri** *Multiple Faculty Positions*

The School of Computer Science and Mathematics at the University of Central Missouri is accepting applications for three tenure-track positions in Computer Science at the rank of Assistant Professor (appointment starts August 2022) and five non-tenure track positions in Computer Science at the rank of Assistant Instructor (appointment starts January 2022). We are looking for faculty excited by the prospect of shaping our school's future and contributing to its sustained excellence.

Required Qualifications for Tenure Track Positions:

- ▶ Ph.D. in Computer Science by August 2022
- ▶ Research expertise and/or industrial experiences in Cloud Computing, Data Science, Operating Systems or Software Engineering
- ▶ Demonstrated ability to teach existing courses at the undergraduate and graduate levels
- ▶ Ability to develop a quality research program and secure external funding
- ▶ Commitment to engage in curricular development/assessment at the undergraduate and graduate levels
- ▶ A strong commitment to excellence in teaching, research, and continued professional growth
- ▶ Excellent verbal and written communication skills

The Application Process for Tenure-Track Positions: To apply online, go to <https://jobs.ucmo.edu>. Apply to positions #998453, #998454 or #998560. The following items should be attached: a letter of interest, a curriculum vitae, a teaching and research statement, copies of transcripts, and a list of at least three professional references including their names, addresses, telephone numbers and email addresses. Official transcripts and three letters of recommendation will be requested for candidates invited for on-

campus interview. Initial screening of applications begins November 15, 2021 and continues until position is filled.

Required Qualifications for Non-Tenure Track Positions:

- ▶ M.S. in Computer Science or a closely related area required; Ph.D. in Computer Science or a closely related areas preferred
- ▶ Demonstrated ability to teach existing courses at the undergraduate and/or graduate levels
- ▶ Excellent verbal and written communication skills

The Application Process for Non-Tenure Track Positions: To apply online, go to <https://jobs.ucmo.edu>. Apply to positions #997336, #997337, #997338, #997344 or #997375. The following items should be attached: a letter of interest, a curriculum vitae, copies of transcripts, and a list of at least three professional references including their names, addresses, telephone numbers and email addresses. Official transcripts and three letters of recommendation will be requested for candidates invited for on-campus interview. Initial screening of applications begins October 15, 2021, and continues until position is filled.

The University of Central Missouri is an Equal Opportunity employer. We accept applications from qualified applicants regardless of race, gender, or abilities. Minorities, women, disabled, and veterans are encouraged to apply.

## **University of Michigan** **Computer Science and Engineering Faculty 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 and across all ranks. Qualifications include an outstanding academic record; an awarded or expected doctorate (or equivalent) in computer science, computer engineering, or a related area. 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 will begin reviewing applications as soon as they are received starting October 1st, 2021 and continuing throughout the year. For more details on these positions and to apply, please visit <https://cse.engine.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 90 world-class faculty members, over 400 graduate students, and a large and illustrious network of alumni. Ann Arbor is known as one of the best small cities in the nation.

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 demonstrated 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: <https://www.engin.umich.edu/culture/diversity-equity-inclusion/>. CSE is firmly committed to DEI and improving our climate through transparent communication and effective action, as shown in our annual report: <https://cse-climate.engin.umich.edu/reports/climate-dei-reports/cse-climate-dei-report-20-21/>.

#### U-M COVID-19 Vaccination Policy

COVID-19 vaccinations are now required for all University of Michigan students, faculty and staff across all three campuses, including Michigan Medicine, by the start of the fall term on August 30, 2021. This includes those working or learning remotely. More information on this policy is available on the Campus Blueprint website.

#### University of Notre Dame Two Faculty Positions

The Department of Computer Science and Engineering at the University of Notre Dame invites applications for two faculty positions. The Department seeks to attract, develop, and retain excellent faculty members with strong records and future promise. The Department is especially interested in candidates who will contribute to the diversity and excellence of the University's academic community through their research, teaching, and service.

One position is a tenure-track position at the Assistant Professor rank in the systems area (IoT, security, etc.). Outstanding candidates in other areas may be reviewed with special consideration for faculty with research interests at the interface of computer science and biology, medicine, and/or health.

The other position is a teaching position at the Assistant Teaching Professor rank.

Applicants must submit a cover letter, a CV, a research statement (if applicable), a teaching

statement, a statement that summarizes their planned contributions to diversity, equity, and inclusion, and contact information for three professional references. To guarantee full consideration, applications must be received by November 5, 2021; however, review of applications will continue until December 17, 2021. Information about all positions may be found at <https://cse.nd.edu/about-cse/faculty-job-openings> including links to the specific job openings.

The Department offers the Ph.D. degree and undergraduate Computer Science and Computer Engineering degrees. Faculty members are expected to excel in classroom teaching and to serve the profession and the University. Tenure track faculty members are expected to lead highly-visible research projects that attract substantial external funding, and to advise graduate students. More information about the department can be found at: <https://cse.nd.edu/>.

The University is an Equal Opportunity and Affirmative Action employer; we strongly encourage applications from women, minorities, veterans, individuals with a disability and those candidates attracted to a university with a Catholic identity.

#### Vanderbilt University Tenure-Track Faculty Positions in Computer Science

The Department of Computer Science (CS) launched in 2020 a multi-year faculty recruitment and hiring process for 20 tenure-track positions at the Assistant, Associate, and Full Professor levels over and above normal hiring patterns, with preference at early-career appointments. In the first year of the initiative, the department welcomed eight new faculty members. In the second year, the initiative will support at least eight new faculty positions starting in the 2022-2023 academic year. Destination Vanderbilt-CS is part of the university's recently launched Destination Vanderbilt, a \$100 million university excellence initiative to recruit new faculty. Over the next three years, the university will leverage the investment to recruit approximately 60 faculty who are leaders and rising stars in their fields. All hires who are part of this initiative are over and above the normal faculty hiring rate at the university.

We seek exceptional candidates in broadly defined areas of computer science that enhance our research strengths in areas that align with the following investment and growth priorities of the Vanderbilt University School of Engineering:

1. Cybersecurity and Resilience
2. Autonomous and Intelligent Human-AI-Machine Systems and Urban Environments
3. Computing and AI for Health, Medicine, and Surgery
4. Design of Next Generation Systems, Structures, Materials, and Manufacturing

Our priorities are designed to ensure the strongest positive impact on computer science and cross-disciplinary areas at all six academic departments in the School of Engineering and other colleges and schools across campus. The hiring initiative builds on these strengths and aspires to propel the Vanderbilt CS Department

#### 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 or a closely affiliated discipline by the time of their appointment (anticipated 1 September 2022).

We are particularly interested in applicants specializing in one or more of the following areas: radar cross section analysis, low observables, electromagnetic scattering analysis, computational electromagnetics, antennas and propagation, or microwave theory and measurements. Applicants having experience in the electromagnetic survivability community are highly desired. This position requires teaching at the graduate level as well as establishing and sustaining a strong Department of Defense 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 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.

For more information on the position and how to apply, please visit <https://www.usajobs.gov/GetJob/ViewDetails/jobadnumber>. Be sure to include

- A letter of application to include the USA Jobs announcement number jobadnumber.
- Your curriculum vitae (no photographs please).
- Transcripts for all degrees listed on curriculum vitae (official copies must follow).
- A statement of your research plans (limited to one page) and a statement of your teaching philosophy at the graduate level (limited to one page).
- A list of three professional references including name, complete mailing address, email address, and phone number.

Applicants must be U.S. citizens and currently hold or be able to obtain a security clearance. More information on AFIT and the Department of Electrical and Computer Engineering can be found at <http://www.afit.edu/ENG/>. Review of applications will begin on January 3, 2022. The United States Air Force is an equal opportunity, affirmative action employer.

to one of the leading academic programs nationally and beyond. Successful candidates are expected to teach at the undergraduate and graduate levels and to develop and grow vigorous programs of externally funded research.

Ranked #14 nationally, Vanderbilt University is a private, internationally recognized research university located on 330 park-like acres 1.5 miles from downtown Nashville, Tennessee. Its 10 distinct schools share a single cohesive campus that values collaboration. The university enrolls over 13,500 undergraduate, graduate, and professional students, including 36% minority students and over 1,100 international students from 84 countries. The School of Engineering is on a strong upward trajectory in national and international stature and prominence, and has built infrastructure to support a significant expansion in faculty size. In the rankings of graduate engineering programs by U.S. News & World Report, the school ranks in the top 20 private, research-extensive engineering schools. Five-year average T/Tk faculty funding in the formerly combined EECS department is above \$800k per year per person. Nearly all junior faculty members hired during the past 15 years have received prestigious young investigator awards, such as NSF CAREER and DARPA CSSG.

With a metro population of over two million people, Nashville's top industries by employment include trade, transportation and utilities; education and health services; professional and business services; government; and leisure and hospitality. Other industries include manufacturing, financial activities, construction, and information. Long known as a hub for health care and music, Nashville is a technology center with a considerable pool of health care, AI, and defense-related jobs available. In recent years, the city has experienced an influx of major office openings by some of the largest global tech companies and prime Silicon Valley startups.

Vanderbilt University has a strong institutional commitment to recruiting and retaining an academically and culturally diverse community of faculty. Minorities, women, individuals with disabilities, and members of other underrepresented groups, in particular, are encouraged to apply. Vanderbilt is an Equal Opportunity/Affirmative Action employer.

Vanderbilt University has made the safety of our students, faculty and staff, and our surrounding communities a top priority. As part of that commitment, the University recently announced that students, faculty, and staff, are required to be vaccinated against COVID. As a prospective and/or a new employee at Vanderbilt, you will be required to comply with the University's vaccination protocol. Effective, August 1, 2021, proof of full vaccination or an approved accommodation will be required before the start of employment in order to work at Vanderbilt University.

Applications should be submitted on-line at: <http://apply.interfolio.com/94225>. For more information, please visit our web site: <http://vu.edu/destination-cs>. Applications will be reviewed on a rolling basis beginning December 1, 2021 with interviews beginning January 1, 2022. For full consideration, application materials must be received by January 31, 2022.



## Faculty Positions in Computer and Communication Sciences

at the Ecole polytechnique fédérale  
de Lausanne (EPFL)

The School of Computer and Communication Sciences (IC) at EPFL invites applications for tenure track faculty positions in *all* areas of computer and communication sciences. The appointments will be at the assistant professor level, but senior appointments are also possible.

Candidates must have an outstanding academic record, a compelling high-impact vision, and a strong commitment to excellence in teaching and mentoring students.

EPFL attracts top students from all over the world and offers competitive salaries, generous research funding and excellent research infrastructure.

Switzerland has an exceptionally high human development index and consistently ranks highly in quality of life, economic competitiveness, and innovation.

Applicants must submit a cover letter, a curriculum vitae including a publication list, brief statements of research and teaching interests, and contact information of at least 3 references (for senior positions, at least 5) who are ready to supply a letter upon request.

Applications must be uploaded in PDF format to the recruitment website:

<https://facultyrecruiting.epfl.ch/position/34865159>

Screening will start on **December 1, 2021**, but applications submitted after this date will also be considered.

Further questions can be addressed to:

**Prof. George Candea**

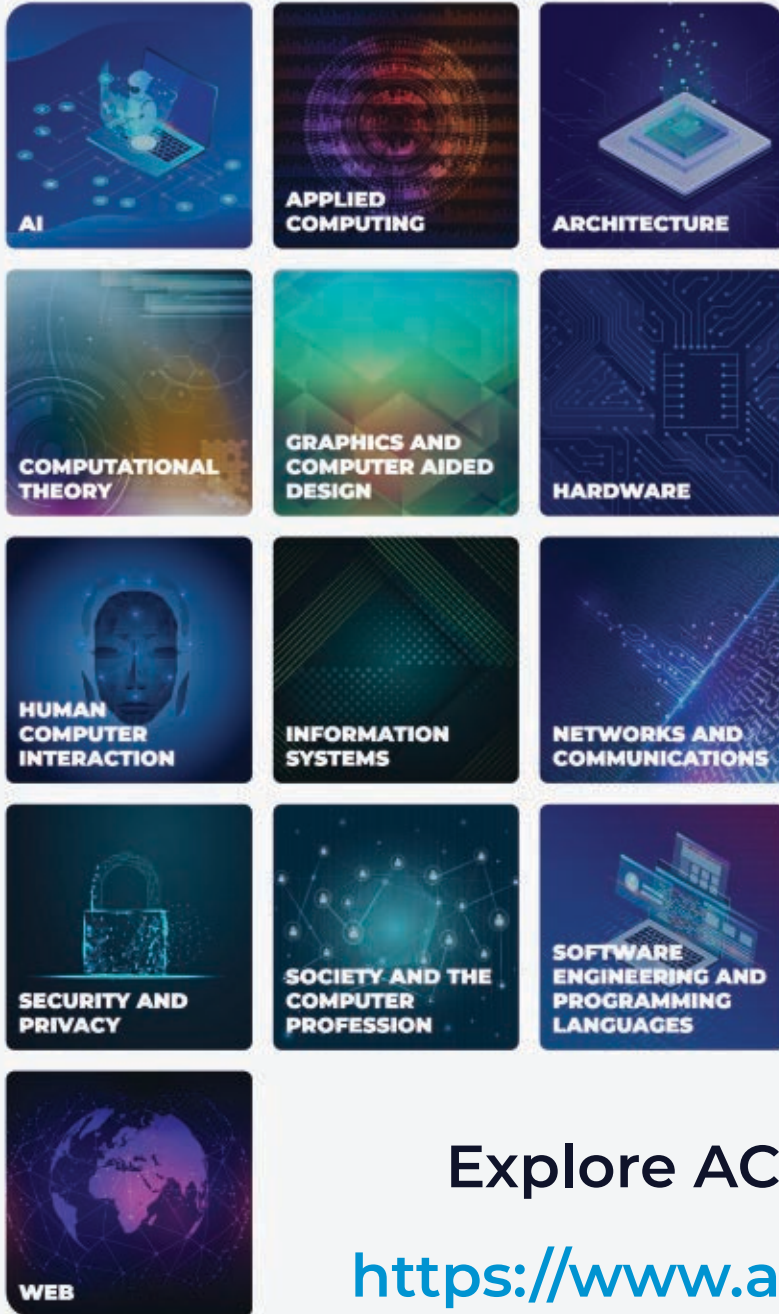
Chair of the Faculty Recruiting Committee

Email : [recruiting.ic@epfl.ch](mailto:recruiting.ic@epfl.ch)

*For more information on EPFL and our School, please visit:  
[www.epfl.ch](http://www.epfl.ch) and <https://ic.epfl.ch>.*

EPFL is an equal opportunity employer and a family friendly university. We are committed to improving the diversity of our faculty.

# Introducing **ACM Focus**



## **A New Way to Experience the Breadth and Variety of ACM Content**

ACM Focus consists of a set of AI-curated custom feeds by subject, each serving up a tailored set of the latest relevant ACM content from papers to blog posts to proceedings to tweets to videos and more. The feeds are built in an automated fashion and are refined as you interact with them.

**Explore ACM Focus today!**

<https://www.acm.org/acm-focus>



Association for  
Computing Machinery





[CONTINUED FROM P. 144] Hobbit mathom house. Just as our raid team vanquished a Hobbit boss named Babbage, I got another clue in the text chat: “You got GE, now look for N in Narrational.”

This amateur-programmed game was turgid in the extreme, lacking graphics and played via email. It consisted of one huge decision tree of If-Thens, each selection taking a verbose, text-based story in a different direction, with no battles or beautiful scenery to provide pleasure. The first decision concerned whether the narrative’s main character would be “good and honest” (select  $A = A$ ) or “bad and deceptive” ( $A = B$ ). I chose the good path, which led nowhere, so I restarted as a bad character and always chose the worse alternative. When I reached the worst possible climax, step Z, I was surprised to earn the symbol “ $\neq$ ” meaning “not equal.” After cursing and meditating, I returned to the very beginning, and instead of selecting  $A = A$ , or  $A = B$ , was now able to choose  $A \neq A$ . I immediately earned:

“You have GEN, now seek I in Infernal.” (I wondered if G-E-N referred to General Semantics, the linguistic philosophy that claimed nothing is itself?)

Oddly, *Infernal* was a dark occult horror show situated in a battle between two fleets of sci-fi spaceships, one operated by a cult having the uninformative name “This,” and the other belonging to the equally irrational “Deji.” One could play solo, as a character unimaginatively named Solo, or stumble across other players in a version where each avatar was assigned a short name composed of random letters, like Duck, Dark, or DARTH. It did not take long to figure out that I needed to learn some mysterious magic spell which would enable travel faster than the speed of light, gravity to be switched on or off inside the spaceship, or a little robot resembling a garbage can to tell jokes that were actually funny. “Academy it is that you seek, Master GENI!” exclaimed a frog whose name was something like Day-Zero, apparently unrelated to any puzzle I had solved and suggesting that chaos ruled this galaxy.

Academy at first looked like a plagiarized version of *Fallout 4*. It was set in some institution of higher learning in Cambridge, Massachusetts, after civilization had collapsed, but I found no other similarities. Avatars looking like college students sat around a table

**I opened the box to see six glorious, perforated Masonite disks, the brass jumpers and wires, batteries, and little light bulbs.**

playing the equivalent of a card game on tablet computers. The system running the game distributed random fragments of code, like cards, but in these four suits: Fortran, Cobol, Pascal, and BASIC. Players had to understand all the code so they could group the fragments to make the best hand. For example, I made a full house when two fragments performed one function, such as Pascal and BASIC code displaying the same bar graph, and three other fragments performed another function, such as erasing a player’s most recent file. That hand earned me the fifth letter, so I had G-E-N-I-A. Remarkably, with that clue, I was able to find the answer on Wikipedia, so I knew what I had to find in the final game: *Computer Museum*.

I searched through the museum’s virtual halls, looking for a small cardboard box hidden among massive hardware junk. There! I opened the box to see six glorious, perforated Masonite disks, the brass jumpers and wires, batteries, and little light bulbs. The most marvelous computer of all time—the educational GENIAC, or “GENIus Almost-automatic Computer,” which I had received for my birthday way back in 1956. Teenagers could assemble many different logical circuits to solve If-Then puzzles, play games, and ignite the sparks that would create the future world of computer science.

**William Sims Bainbridge** ([wsbainbridge@hotmail.com](mailto: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.

© 2021 ACM 0001-0782/21/11 \$15.00

## Distinguished Speakers Program

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

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

**[speakers.acm.org](http://speakers.acm.org)**



Association for  
Computing Machinery

From the intersection of computational science and technological speculation, with boundaries limited only by our ability to imagine what could be.

DOI:10.1145/3484694

William Sims Bainbridge

## Future Tense World of Hackcraft

*An obsessive gamer's quest for the absolutely most significant computer ever.*

LET ME INTRODUCE myself as who I really am rather than role-playing through one of my hundred avatars. I am Leigh Jenkins, a professional journalist for the gamer blog *Highly Ludic*, who mainly writes about developments in massively multiplayer online role-playing games, aka MMORPGs. My writing career began in 1997, when I first explored *Ultima Online*, which still exists nearly a quarter-century later, while many of the best subsequent games—including *The Matrix Online*, *Tabula Rasa*, and *Fallen Earth*—have gone out of business.

During the first few months of 2021, I had been storming through illegal rogue servers, looking for MMORPGs that had been shut down by the companies that owned them, such as *Star Wars Galaxies* and *City of Heroes*, cancelled in 2011 and 2012 respectively. Then I learned that the *Defiance* skirmish MMO, which had originally launched in 2013, was shutting down at the end of April, so I asked my readers if any of them planned to set up a rogue version.

Several of my adorable fans told me I was outdated; I should know that law-abiding hackers created an MMORPG-MWM (MWM = Multi-World Maze) named Hackcraft—a virtual-reality version of the antique social media webring structure that predated Facebook. Unlike the old-fashioned rogue games, Hackcraft was legal because it assembled metaphors based on dozens of games rather than duplicating one and violating copyrights.

In any of Hackcraft's subgames, a user accepts a major quest arc, the completion of which opens the door to the next subgame on the ring. When



I entered the hub, I needed to select a mystery to solve, and I chose:

“What is the name of the absolutely most significant computer ever created?”

The tutorial told me to seek the letter “G” in the quest arc of a game called *Guild of the Rings*, but it was not immediately obvious what that meant. Clicking the “GO” button hurled my avatar into the game, specifically into The Prancing Pony, a tavern where drunken avatars were singing,

*“Far over the misty mountains grim,  
to dungeons deep and caverns dim.”*

Starting at Level 1, a player ventures into the dangerous world outside the tavern to collect resources, such as hops and barley to brew beer and other essential components of tavern life. Unfortunately, each resource is in a different location, guarded by increasingly powerful non-player enemies. Players with-

in a guild form “bands,” not for fighting but for playing music (using the ABC scripting language) and for dancing like a prancing pony. I deduced that the way to find the “G” was to create virtual objects, so I developed crafting skills and hunted for the recipes required to make simple computers. I began with a sundial, discovering that the part that casts the shadow is a Gnomon (the “G” is not even pronounced). That was a sophisticated but false hypothesis. My second guess turned out to be correct: the letter being sought was not really a clue; upon building a “slipstick” or slide rule, I triggered this message: “You got G, now look for E in Elfdom.”

*Elfdom*, a subgame in the amateur ring, had millions of players but a very narrow fantasy backstory. Elf Doom would have been a more appropriate name. In creating your avatar, you select from one of the long lists of races which are totally hostile to one another but are merely different varieties of elves. It is perfectly clear why the nature-loving Night Elves warred with the technocratic Blood Elves. But I never could fathom why the Dark Elves, High Elves, Half Elves, Wood Elves, Eth Elves, Grey Elves, Painted Elves, and Deep Elves hated each other so much. You would think that the 10 varieties of Elves, all played by people, would unite against the two computer-controlled enemy races, the Halflings and Hobbits. Each quest arc must be completed by a team uniformly comprising one type of elf—sneaking deep into enemy territory and often battling other elf teams to assassinate an enemy boss in a Halfling temple or [CONTINUED ON P. 143]

# IEEE BITS

THE INFORMATION THEORY MAGAZINE

**A FEW BITS OF INFORMATION CAN  
MAKE ALL THE DIFFERENCE!**

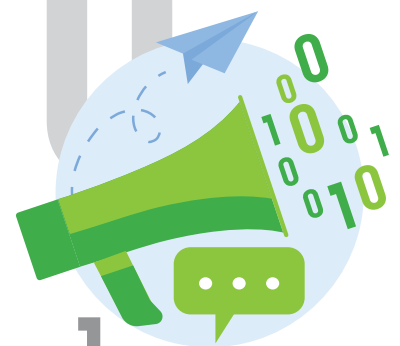
**BITS** brings you tutorial articles and columns exploring topics such as quantum information and computing, blockchain, information inequalities, privacy, imaging and machine learning through an information theoretic lens.

Included in print and electronic format with an IEEE Information Theory Society Membership (\$25 for Regular Members, \$5 for Student Members in 2022).

We welcome contributions that infuse new problems into information theory and diffuse information theoretic thinking into emerging areas.

Please submit at:  
<https://mc.manuscriptcentral.com/bits-ieee>

<https://www.itsoc.org/bits>





# MATLAB SPEAKS MACHINE LEARNING

With MATLAB® you can use clustering, regression, classification, and deep learning to build predictive models and put them into production.

[mathworks.com/machinelearning](https://mathworks.com/machinelearning)

©2021 The MathWorks, Inc.