SPECIAL ISSUE ON THE IMPACT STATISTICIANS AND DATA SCIENTISTS HAVE IN THEIR COMMUNITIES

September 2022 • Issue #543

AMSTATNEWS

The Membership Magazine of the American Statistical Association • *http://magazine.amstat.org*

DATA LITERACY as a TOOL for COMMUNITY HEALTH *and* SOCIAL JUSTICE

ALSO:

Prescribing Privacy: Human and Computational Resource Limitations

Statisticians and Wildfires at the Wildland/ Urban Interface

JSM 2022 WASHINGTON, DC

Thank you to our JSM 2022 sponsors

PLATINUM

abbvie

sas

Takeda

GOLD

AstraZeneca

BeiGene

Berry Consultants Statistical Innovation

Lilly

jmp STATISTICAL DISCOVERY

MERCK INVENTING FOR LIFE

Meta

NSA WHERE INTELLIGENCE GOES TO WORK®
www.IntelligenceCareers.gov/NSA

STATA

TWO SIGMA

UNLEARN

Westat

SILVER

ADDINSOFT

Biogen

Bristol Myers Squibb

Carnegie Mellon University Software Engineering Institute

Daiichi-Sankyo

MIT MANAGEMENT

NSF

NOVARTIS

Otsuka

Pacific Northwest National Laboratory

P&G

RTI INTERNATIONAL

United States Census Bureau

VERTEX

WILEY

AMSTATNEWS

SEPTEMBER 2022 • ISSUE #543

Amstat News welcomes news items and letters from readers on matters of interest to the association and the profession. Address correspondence to Managing Editor, *Amstat News,* American Statistical Association, 732 North Washington Street, Alexandria VA 22314-1943 USA, or email *amstat@amstat.org*. Items must be received by the first day of the preceding month to ensure appearance in the next issue (for example, June 1 for the July issue). Material can be sent as a Microsoft Word document, PDF, or within an email. Articles will be edited for space. Accompanying artwork will be accepted in graphics file formats only (.jpg, etc.), minimum 300 dpi. No material in WordPerfect will be accepted.

Amstat News (ISSN 0163-9617) is published monthly by the American Statistical Association, 732 North Washington Street, Alexandria VA 22314-1943 USA. **Periodicals postage paid** at Alexandria, Virginia, and additional mailing offices. POSTMASTER: Send address changes to *Amstat News,* 732 North Washington Street, Alexandria VA 22314-1943 USA. Send Canadian address changes to APC, PO Box 503, RPO West Beaver Creek, Rich Hill, ON L4B 4R6. Annual subscriptions are $50 per year for nonmembers. *Amstat News* is the member publication of the ASA. For annual membership rates, see *www.amstat.org/join* or contact ASA Member Services at (888) 231-3473.

American Statistical Association
732 North Washington Street
Alexandria, VA 22314–1943 USA
(703) 684–1221
ASA GENERAL: *asainfo@amstat.org*
ADDRESS CHANGES: *addresschange@amstat.org*
AMSTAT EDITORIAL: *amstat@amstat.org*
ADVERTISING: *advertise@amstat.org*
WEBSITE: *http://magazine.amstat.org*

The American Statistical Association is the world's largest community of statisticians. The ASA supports excellence in the development, application, and dissemination of statistical science through meetings, publications, membership services, education, accreditation, and advocacy. Our members serve in industry, government, and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare.

FEATURES

LEADERS IN COMMUNITY ANALYTICS

COMMUNITY ANALYTICS

A Date with Data:
Stepping Toward Data Literacy

Page 22

Building Statistics and Data Science Capacity for Development

The LISA 2020 Network began almost a decade ago with the idea that collaborative statisticians in developing countries could create stat labs to build statistics and data science capacity. Based on the collective experiences of more than 30 newly created stat labs since then, the network is being transformed by the idea of building statistics and data science capacity by focusing research, education, and outreach efforts on the intersections of data-driven development. Read more about this at *https://magazine.amstat.org*.

FDA to Hold Public Workshop to Discuss Biosimilar Product Development

The US Food and Drug Administration is hosting a one-day virtual public workshop called "Increasing the Efficiency of Biosimilar Development Programs" on September 19 from 9 a.m. to 4 p.m.

This public workshop will focus on statistical, scientific, and clinical methods for streamlining comparative clinical studies associated with biosimilar product development programs.

The workshop is open to the public; however, registration is required at *https://bit.ly/3pcXiJA*.

For details, visit *https://bit.ly/3payMIX*.

In Memoriam

Sadly, two prominent statisticians passed away recently. Distinguished research professor and professor of statistics from The George Washington University **Nozer D. Singpurwalla** died on July 22 at his home in Washington, DC. **William (Bill) Winkler,** longtime ASA member and statistician for the US Census Bureau passed away on June 30. View their obituaries at *https://magazine.amstat.org*.

Correction

In the August issue of *Amstat News,* the month registration opens for The Curiosity Cup Challenge was incorrect. Registration begins in October. We apologize for the error.

COLUMNS

DEPARTMENTS

MEMBER NEWS

Follow us on Twitter
www.twitter.com/AmstatNews

Join the ASA Community
http://community.amstat.org

Like us on Facebook
www.facebook.com/AmstatNews

Follow us on Instagram
www.instagram.com/AmstatNews

Subscribe to our YouTube channel
www.youtube.com/user/AmstatVideos

JSM 2022 Foundations and Innovations and Looking Forward

A Foundation for Innovation was the theme for JSM 2022. When I selected it, I wanted to celebrate our contributions to science and society. My JSM 2022 experiences reminded me that a most important foundation is our community. I won't be able to capture in words the excitement of seeing old friends, the serendipity of making new friends during a technical session, or the joy of being able to celebrate our many successes, so I need your help. Please share the "selfies" you took to commemorate JSM using the form at *https://form.jotform.com/zzlalo/jsm-2022-selfies.* I'm including a few of mine here!

For members of the ASA Board of Directors, JSM began on Friday, August 5, with the first day of our two-day board meeting. The agenda was packed, and a more detailed overview will be published in a subsequent *Amstat News* article, so I want to focus on one outcome. The board voted to establish a new outreach group: the Caucus of Industry Representatives.

This group will parallel the successful Caucus of Academic Representatives. Its purpose is to promote statistics and data science in the private and public sectors and provide resources for industry statisticians and data scientists to successfully advocate for the discipline. I am grateful to Ginger Holt, senior staff data scientist for Databricks, for agreeing to lead this effort. If you are interested in learning more about opportunities to be involved, reach out to ASA Director of Strategic Initiatives and Outreach Donna LaLonde at *donnal@amstat.org*.

Planning my conference schedule was different this year because business meetings took priority over the technical program. I was fortunate to be able to meet with the Caucus of Academic Representatives, Leadership Council, and Membership Council, as well as attend the COPSS meeting, the session recapping the findings and recommendations of the anti-racism task force, and many more important business meetings of our association. I left each meeting amazed by the breadth of work and contributions from members. If you are not involved in a chapter or section, I encourage you to get involved. It will be rewarding.

The program committee's hard work paid off. Led by Ming-Hui Chen, the committee selected the introductory overview lectures and late-breaking sessions, which included presentations about computational advertising, sports analytics, transparency in federal statistics, and algorithmic bias and public policy.

Linda Young (left) and Katherine Ensor stop for a photo before Ensor gives the ASA President's Address.

Katherine Ensor

Monnie McGee (left) and Steve Sain flank Katherine Ensor in a selfie while celebrating during JSM.

I will admit to FOMO (fear of missing out) when it comes to the technical program, since my schedule did not allow me to attend many sessions. However, I know from conversations with colleagues and Twitter comments that the quality of sessions once again highlighted the breadth and depth of our science. I think Maria Cuellar's tweet (*https://bit.ly/3T0DuGX*) captures the celebratory spirit. She says, "I loved being a part of two sessions in statistics and the law. The other talks were so much fun to listen to. Looking forward to the dance party to celebrate. See you there!"

Rob Santos (right) tweeted, "So happy to see upcoming 2023 American Statistical Association President Dionne Price at #JSM2022."

Word clouds made from audience members' responses to Katherine Ensor's request to share their idea of "innovations" and "foundations"

In the past, one of my first steps in JSM preparation would be to add the featured speakers to my schedule. Looking back, it is worth celebrating the rich history of these plenary talks. How many of you were in the audience in 2011 when Sir David Cox was one of the featured speakers (*www.youtube.com/watch?v=LE3rhuD7zhk)*?

This year, I had the pleasure of introducing Reginald DesRoches as my invited speaker and David Banks as the Deming lecturer. We also had Medallion lectures by Dylan Small and Huixia Judy Wang. Madhu Mazumdar gave an inspirational COPSS Elizabeth L. Scott Lecture, and Nancy Reid was recognized with the COPSS Distinguished Achievement Award. As Brahmar Mukherjee shared on Twitter (*https://bit.ly/3c52HPX)*, Reid was the first woman to receive the COPSS Presidents' Award 30 years ago. She is indeed a statistical hero! This year, we recognized Daniella Whitten, a true leader in our field, with the COPSS Presidents' Award. Remember, all the plenary talks will be publicly available on the JSM 2022 website.

The ASA vision imagines a world that relies on data and statistical thinking to drive discovery and inform decisions. Since JSM was in the US capital this year, we had a unique opportunity to promote the importance of data science and literacy. Members of our community participated in Capitol Hill visits to advocate for a data science and literacy bill. The bill would create a voluntary program, which would be administered by the US Department of Education, to increase K–16 student access to data science and data literacy educational opportunities. Educational institutions would be able to apply for grant support of programs that include professional development for teachers, workforce development, and curriculum development. A big thank you to those who added Hill visits to their JSM schedule! This effort is ongoing, so to get involved, contact Steve Pierson, ASA director of science policy, at *spierson@amstat.org*.

My presidential address, titled "Statistical Foundations Driving 21st-Century Innovation," brought forward the importance of our science. I asked you to share your idea of "innovations" and "foundations" during the talk and, for fun, I include the resulting word cloud in this article. There is still time to grow our collective voice. If you would like to add to the conversation, visit *www.menti.com/h49s3fzbjs* for innovations and *www.menti.com/ddcvo5fw1r* for foundations. I will speak to this further in the *JASA* article associated with my presidential address.

To be able to share innovations and foundations with you and to recognize the ASA award winners and fellows was an experience I will treasure. Seeing the videos of those we were celebrating let us learn more about them. The heartfelt sentiment from the conclusion of my speech remains true: What you do matters!

My ASA Story:

Maria Tackett, Assistant Professor

. . . there is a common goal to continue researching and innovating to provide effective and engaging learning experiences for all students.

I'm delighted to share my ASA story about my involvement with the Section on Statistics and Data Science Education.

I joined the ASA in 2015 as a graduate student, but my involvement with SSDSE really began when I attended my first US Conference on Teaching Statistics at Penn State in May 2019. At that point, I had just completed my first academic year as a faculty member in the department of statistical science at Duke University and was looking forward to meeting more statistics educators and getting involved in the statistics education community.

The opening session had a series of five-minute talks, a tradition at USCOTS and its electronic counterpart, the Electronic Conference on Teaching Statistics. New and interesting ideas were presented in each short presentation, so I knew I was going to learn a lot during the conference.

I had the opportunity to present a poster about my undergraduate regression course; it was the first time I presented work in statistics pedagogy. Though I was nervous at the start of the session, I was quickly energized by the feedback, encouragement, and exchange of ideas.

These engaging conversations continued throughout the conference as I attended breakout sessions and discussions about a variety of topics, ranging from new pedagogies and innovative classroom technology to facilitating constructive classroom conversations on challenging topics. These types of discussions didn't end at USCOTS, as SSDSE members regularly connect through events such as a recent webinar about mentoring undergraduate research and the section discussion forum.

Over the past three years, I have become more involved in SSDSE and am currently the communications chair for the section. In this role, I maintain the section blog with news, updates, and articles about topics of interest to members. I also implemented a new feature called "Meet a Member," which periodically spotlights an SSDSE member.

In addition, I participate in the section's mentorship program that pairs early-career educators with experienced educators and statistics education researchers. I was paired with Beth Chance from California Polytechnic State University in the most recent academic year. She mentored me as I wrote a paper about the pedagogy of a modern undergraduate regression course based on my experience teaching the same course I presented at my first USCOTS.

Through these and other experiences in SSDSE, I have been fortunate to meet statistics educators from K–12 and higher education institutions around the country. Though we teach different student populations, there is a common goal to continue researching and innovating to provide effective and engaging learning experiences for all students. I have been inspired by the members of SSDSE, and I look forward to being part of this community for years to come. ■

Stephanie Shipp: 'Democratizing' Data Science to Serve the Public Good

Kim Gilliam, ASA Marketing and Communications Coordinator

Stephanie Shipp

Stephanie Shipp's Arlington, Virginia, office overlooks the Potomac River and offers a sweeping view of the Washington Monument and other familiar landmarks of Washington, DC—a town she knows well from her days at the Federal Reserve Board, Bureau of Labor Statistics, US Census Bureau, and National Institute of Standards and Technology.

Today, Shipp is the interim director and a professor at the Social and Decision Analytics Division (SDAD) within the Biocomplexity Institute at the University of Virginia. Working with Sallie Keller, the founding director, she built and developed the division, which was purposely located in the DC metropolitan area because of its proximity to local, state, and federal policymakers. Shipp and Keller's vision of "democratizing" data science to serve the public good fuels their mission: To provide evidence-based insights to change the way communities make policy to improve lives at the local level.

"When we talk about community analytics, we primarily work with local government officials," Shipp says. "They are passionate about their work and making sure that they're meeting the needs of all people, not just a few, especially the more vulnerable populations. Local officials truly care about their constituents. You hear this concern in every conversation we have."

Connecting the Data

"Local officials are eager to make data-informed decisions, but they often don't have the resources to do that," Shipp says. She recalls working with the Arlington County fire chief, who sought to use his data to improve his "situational awareness" and make better decisions. For example, one of his concerns was to ensure their limited number of medical units were in the right place at the right time of day. He had several silos of data from such places as the call center, where they deployed their units, and after-action reports.

"He had a lot of data, but none of it was connected," Shipp says.

"We undertook the process to statistically link these data by time and geography, and then connected the fire/EMS data to the American Community Survey and social media," says Shipp. "We provided the fire chief a corpus of data, as well as data insights about incidents by season, when special events occurred, and by neighborhood. The culmination of this work provided valuable insight to support local and future decision-making."

Community Learning for Data-Driven Discovery: Modeling Success

Part of this successful partnership is based on a research model developed by Keller, Shipp, and the research division's team—the Community Learning Through Data-Driven Discovery, or CLD3 for short.

Shipp explains the innovative approach. "At its core, CLD3 provides researchers with a guide and the philosophy that defines how we begin our work with a community. We don't go into a community and say, 'We want to study X.' We approach them with, 'What are your challenges? What keeps you up at night? What problems can't you solve?' But it doesn't stop there. We continue to engage with community officials throughout the entire process. We brainstorm and identify data sources, context, and results. We work together as partners.

"Partnering is critical because we are not just handing them a report with the results that leaves them wondering, 'Well, how do I use this?' They're involved at every step of the way. It's incredibly rewarding."

> We're talking with local government leaders and policymakers who are extremely interested in equity.

Stephanie Shipp heads to the National Gallery of Art via Metro bus and train with her grandkids, Myrna, age 5, and Francis, age 7.

Scaling the Mission

The Social and Decision Analytics team's vision is to scale the CLD3 process nationally. They developed a strategy to bring the CLD3 process to local governments everywhere by leveraging the expertise of land-grant universities that disseminate information to communities and rural areas. The plan capitalizes on the expertise of the Cooperative Extension System. "Extension has been around for over 100 years and is the boots on the ground—they know their communities well," Shipp explains.

In 2020, a three-state team received funding from the US Department of Agriculture and Bill and Melinda Gates Foundation to conduct a pilot to test their strategy across three states: Iowa, Virginia, and Oregon. They focused on advancing economic mobility by working with extension professionals on data-driven projects in their communities.

One project in Virginia was to look at access to health care when the community's only hospital closed. The team received additional funding to extend this work across all counties in Virginia. Leaders in another community wanted to know how to successfully transition people who've been in jail back into society.

In Iowa, a CLD3 team identified communities in greatest need of excessive alcohol prevention resources. Another team expanded and enhanced the Iowa State University Extension Community Helpline services across the state by developing data science tools to capture customer service, monitor success, and auto-generate reports. This allowed helpline workers to spend more time working with residents and less time filling out paperwork.

In Oregon, one of the CLD3 teams examined the impact of regulations on economic development in eastern Oregon. Another team created an economic mobility baseline for the South County Wasco area.

The Tip of the Iceberg

Currently, the Social and Decision Analytics team is partnering with Mastercard's Center for Inclusive Growth to build a social impact data commons for the Washington, DC, metropolitan area. They are working with local governments in the region to create the data commons, a knowledge repository designed to answer questions that matter most to community leaders.

"We're talking with local government leaders and policymakers who are extremely interested in equity. For example, does everyone in their community have equitable broadband access, and can they afford it? Do they have equitable access to food that meets their cultural needs?" explains Shipp.

They are also collaborating with the Virginia Department of Health, which publishes a large amount of data. The department would like their data to be more easily accessible to increase understanding of their metrics, initially their rural health priority metrics. The data commons is a tool various audiences can use to explore data insights in their regions, communities, or neighborhoods.

"We're taking their data, adding to it, and presenting it in new and interesting ways to help leaders answer questions that keep them up at night," says Shipp. "We are developing these data commons to be replicable by other metropolitan areas and states to set the stage to begin scaling our work nationally. We've only seen the tip of the iceberg." ■

Xihong Lin: On the Front Lines of COVID-19 Research

Kim Gilliam, ASA Marketing and Communications Coordinator

Xihong Lin

In January 2020, the World Health Organization announced a mysterious coronavirus-related pneumonia in Wuhan, China, the capital and largest city in Hubei Province. Very quickly, Wuhan made the extraordinary move to shut down access to the city and isolate its population of 11 million from the rest of the country.

Xihong Lin, biostatistics and statistics professor at Harvard University, became concerned about her former post-doctoral fellow, Chaolong Wang, who was on faculty at the Tongji School of Public Health at the Huazhong University of Science and Technology in Wuhan. She texted to check in on him and his family.

MORE ONLINE
View additional information and links related to this article on *Amstat News* online: *https://magazine.amstat.org.*

Wang shared that he and a colleague were analyzing the Wuhan COVID-19 data. Lin says, "We already had a case in Seattle and one in Boston, so intuition told me this disease might spread and I decided to join them on this research. As spread can happen very quickly, one must react very quickly, otherwise—as we know—it will be hard to control."

The research was a steep learning curve for Lin and Wang because they did not have an infectious disease background, but people who did were part of the team. "We were learning on the fly, but we were determined because we could see this becoming a public health emergency that could have a devastating impact on the world. The team members worked day and night with the goal of sharing the findings to help the world understand what happened in Wuhan and the steps it took to control the outbreak," says Lin.

A preprint of their work was posted on MedRxiv on March 8, 2020, and distributed worldwide. It was later published in *JAMA* and *Nature* and featured in a "behind-the-paper" story in *Nature Portfolio.*

Educating Harvard and Beyond on COVID-19

By the end of February, Lin had a meeting with Harvard leadership. She shared the major epidemiological findings of the research and Wuhan's public health intervention approaches to fighting COVID-19. "At the time, Harvard was contemplating what do with its students considering the COVID threat. I was glad that our Wuhan study findings were helpful to the university in these critical early days. Harvard was among the first to send students home," says Lin.

On March 13, the last day before Harvard went remote, Lin gave a Zoom webinar to the Harvard community about the Wuhan study findings. "I emphasized that, without public health interventions, COVID-19 was very transmissible. I presented the containment steps Wuhan took—from the city lockdown to quarantine and isolation. I also highlighted the extensive personal protective equipment (PPE) used by medical personnel in Wuhan, which included face shields, and indicated that the PPE used by US health professionals was insufficient."

Lin's slides were widely distributed that weekend and, by Monday, there was a national campaign for PPE by physicians. "There was no COVID PPE protocol in place for physicians, and there was not enough PPE to meet their needs," says Lin. "Many physicians reached out to me and were worried, as they were not required to wear masks and did not know what type of PPE was appropriate."

A postdoc in Lin's lab translated the PPE guidelines used in Wuhan from Chinese to English, and her new physician friends edited and shared them with their community. "Back then, physicians didn't know what to wear—a surgical mask or the N-95," Lin says.

"We now know. It was amazing to see how a little statistical talk could spur this kind of effort to help the community."

Lin soon had the media reaching out. She received many invitations for interviews from major media outlets on the Wuhan study findings and strategies to battle the pandemic. "One of the challenges was how to effectively communicate our findings to the general public and policymakers. I hope the ASA can provide media training for statisticians in the future, as public communication skills are very important," says Lin.

She clearly remembers a BBC Radio 4 interview in which she was told by the host before going on the air that the people listening to the show were families with children eating breakfast. She also testified before the UK Parliament's Science and Technology Committee on COVID-19 response in April 2020. "I learned that I needed to convey what the data showed about COVID but communicate without jargon, says Lin. "I had to keep the message simple, short, and focused so the audience can understand in two minutes."

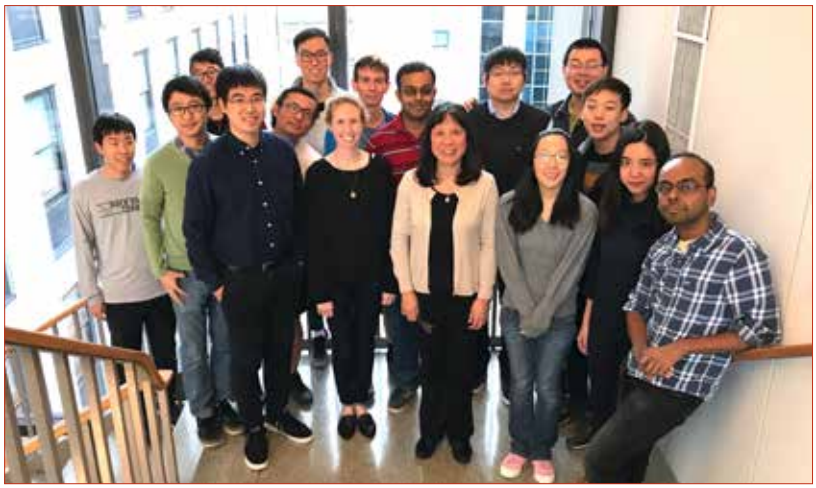
Lin also asked to serve on the COVID-19 Massachusetts State Task Force. She worked with the members to develop recommendations for Gov. Charlie Baker on testing, isolation, quarantine, and contact tracing, in which the state invested a significant amount of money. She also worked with the Broad Institute of MIT and Harvard, a biomedical and genomic research center, on COVID testing. Testing capacity was extremely low then and, working together, the Broad took COVID testing capacity from zero to more than 35 million people by July 2022.

"It was hard to believe that a shortage of testing swabs was a significant problem in spring 2020," Lin says. She found herself taking a masterclass in logistics. Working in tandem with various people, businesses, and the state of Massachusetts, swabs and PPE were shipped their way after passenger planes were converted into cargo flights from China. "It was truly a team effort, and everyone went out of their way to make it happen," recalls Lin.

'Let the Data Speak' Is Not Enough …

One of the big challenges was global implementation of the containment strategies. "The implementation science is critical," says Lin. "Based on the research, we knew the effective public health intervention strategy that works, and the WHO and many public health researchers were onboard, but the problem is many of the containment strategies are not one-size-fits-all. Implementation can vary from one country to another and one culture to another."

"The data science is important, but it's not enough," says Lin. "People's behaviors are very difficult to change. Even if you have the data that shows the intervention works, including something as simple as wearing a mask. Implementing them in the real word is hard.



Xihong Lin is flanked by her postdoctoral fellows and student mentees at the Harvard T.H. Chan School of Public Health in 2018.

"If biostatisticians and statisticians want to make an impact in the world, we must learn how to effectively communicate with the policymakers and the general public on scientific findings, build public trust, and engage them in implementing these recommendations. We also need to work with all the stakeholders to ensure implementations are country and region specific. We can have the most beautiful data findings, but if they cannot be implemented, it will be impossible to have an impact."

Lin's postdoctoral fellows Corbin Quick and Rounak Dey later took a lead in developing and applying epidemic dynamic modeling methods for analysis of the US COVID-19 data. Their work was published in *JASA* as a discussion paper. They found the US COVID-19 data had many complications.

Lin sees the need for improvements in data collection and reporting infrastructure. Public health research relies on access to good data and the ability for research teams to share data and collaborate. She says, "If we don't invest in public health infrastructure, we suffer and the price we pay is big."

Although the last few years have been challenging for Lin, the collaborative work has been incredibly gratifying. "Everyone was a volunteer, and so many jumped in without thinking about receiving credit. They just wanted to help. There was an amazing community spirt," says Lin.

She adds she was excited by how many people in the general public became interested in science during the pandemic. When she posted her preprint on the Wuhan study on Twitter, she gained thousands of followers in a brief time. She also earned the blue verified badge on the platform, which lets people know an account of public interest is authentic, notable, and active. Be sure to follow Xihong Lin at @XihongLin. ■

DATA LITERACY as a TOOL for COMMUNITY HEALTH and SOCIAL JUSTICE

Melody S. Goodman and Janice Johnson Dias

Data is one of the world's most valuable resources and an important tool for improving community health. Therefore, knowledge of how to understand, interpret, and share data is necessary for personal and professional success, most importantly for the work of change-making and improving well-being of low-income and/or predominately minority communities.

In many instances, racialized minority and/or low-income communities are systematically left out of data training and key data collection initiatives, resulting in their underrepresentation both in the field of data science and in many data sources used to make policy decisions. This absence from the data learning and gathering processes has far-reaching public health, political, social, and economic consequences, because we know that even in cases in which there is no or limited available data, there are still important health issues to solve. Inclusion of these populations in data training, collection, and use for social action are pivotal to community well-being.

Understanding these realities, we have spent the last decade working collaboratively with community members to train and find innovative, time bound, relevant solutions to some of the most intractable health issues facing communities. More specifically, though much of the general public health practices rely on evidence compiled by research scientists and other formal health care professionals, we have been laying the foundation for new forms of knowledge generation.

We have been training and equipping community members with the language and skills to understand, interpret, design, collect, and share data. We are working hard to reframe the current perspectives that suggest racialized minority or low-income groups are "hard to reach populations"; we

instead suggest these groups and communities be understood as those "typically excluded using existing approaches."

By focusing on increasing research and data literacy, allowing those most affected by health problems to be part of the development and testing of potential solutions, designing studies that fit in real-world settings with high potential for sustainability, developing meaningful measures for key outcomes, and interpreting research findings in appropriate community context, we are being intentional with the application of our skills. For us, as Black individuals with doctorates who have been educated and raised in largely immigrant, low-income, and urban communities, we are intimately aware of the ways in which knowledge and skill gaps can affect data interpretation, use of services, advocacy, and—most importantly—well-being.

The foundations of this work began with the Community Alliance for Research Empowering Social Change program in Long Island in 2009 and extended to the Community Research Fellows Training program in St. Louis, Missouri, (2012–) and Jackson, Mississippi (2014–).

Goodman has gathered an interdisciplinary team of researchers, community-based organizations, and community members to examine the health issues that disproportionately affect minority populations; developed and implemented culturally competent evidence-based interventions; and disseminated those findings to stakeholders—with the goal of improving minority health and eliminating racial/ethnic health disparities.

In 2018, we adapted the Community Research Fellows Training (for adults) to create the Youth Research Fellows Training (for middle and high school). This was a collaborative effort between New York University School of Global Public Health and the GrassROOTS Community Foundation—a public health and social justice organization that works primarily with Black women and girls.

We trained Black middle- and high-school girls who were participants in the GrassROOTS summer leadership program to become research and data literate partners with social scientists and policymakers so they can advocate for the needs of their communities and create social action campaigns around menstrual equity, green space, and mental health. Participants learned how to develop research questions and were introduced to basic research terminologies, methods, and design.

We were generally satisfied with these efforts, but things changed in the spring of 2020. The COVID-19 pandemic and racial reckoning in the United States underscored the importance of this work and pushed us to deepen our commitment to increasing these proficiencies. The result was the Quantitative Public Health Data Literacy Training (2020–).

Melody S. Goodman, @goodmanthebrain, is the associate dean for research and a professor of biostatistics at New York University School of Global Public Health.

Janice Johnson Dias, @drjanicejohnson, is an associate professor of sociology at John Jay College of Criminal Justice and the president of the GrassROOTS Community Foundation.

Further Reading

The World's Most Valuable Resource Is No Longer Oil, but Data
https://econ.st/3uYmjeD

Training Community Members in Public Health Research: Development and Implementation of Community Participatory Research Pilot Project
https://bit.ly/3PIubcc

Data Literacy
https://en.wikipedia.org/wiki/Data_literacy

A Data and Analytics Leader's Guide to Data Literacy
https://gtnr.it/3yTYfLf

Determining Data Information Literacy Needs: A Study of Students and Research Faculty
http://docs.lib.purdue.edu/lib_fsdocs/23

The Role of Relatedness in Student Learning Experiences
https://eric.ed.gov/?id=EJ1267383

Ethical Guidelines for Statistical Practice
https://bit.ly/3aWkCrC

Too often, community members feel ill-equipped to challenge data findings when presented in the form of tables, graphs, figures, and maps.

The goal of this course is to increase data literacy and improve public health knowledge for youth and adults so they can make sense of the daily bombardment of quantitative information being shared by media sources and government entities. Additionally, we wanted to equip students with the ability to clearly articulate to their families and friends the meaning of the data.

We were seeking to enhance the data literacy of the populations who were likely to be experiencing the negative effects of the public health crisis and the racial onslaught. We wanted desperately to help these community members become data literate—have the ability to understand the contexts that produced the data they were seeing, be able to read and communicate data as information, and have the knowledge to gather and collect data.

Data literacy, therefore, involves understanding what data means, including how to read charts appropriately, draw correct conclusions from data, and recognize when data is being used in misleading or inappropriate ways. Data literacy requires an understanding of constructs, context, data sources, analytical methods, and techniques applied.

The number of applications demonstrates the clear need for data literacy skills. Over the past three years, more than 2,000 individuals have applied to the Quantitative Public Health Data Literacy Training and more than 500 have been trained. Qualitative findings show that these students are excited to learn the information, but are also eager to learn from Black women scholars. Moreover, many report they are interested in filling gaps in their understanding of data.

Too often, community members feel ill-equipped to challenge data findings when presented in the form of tables, graphs, figures, and maps. There are public health consequences for feeling un or underprepared to challenge the stories told about their communities. We provide a space for these students to feel a sense of relatedness, which creates an environment for enhanced learning.

Educational research tells us students learn best when they feel their instructors understand their stories and are committed to their success. This approach to teaching data literacy has helped us communicate the value of being able to read, write, interpret, and present with numbers.

Allowing community members to tell their own stories is key for advocacy and equity. We pay particular attention to quantitative data; however, there remains a need for understanding the people behind the data. Having stories (qualitative data) that support quantitative findings humanizes the numbers and often provides greater impact.

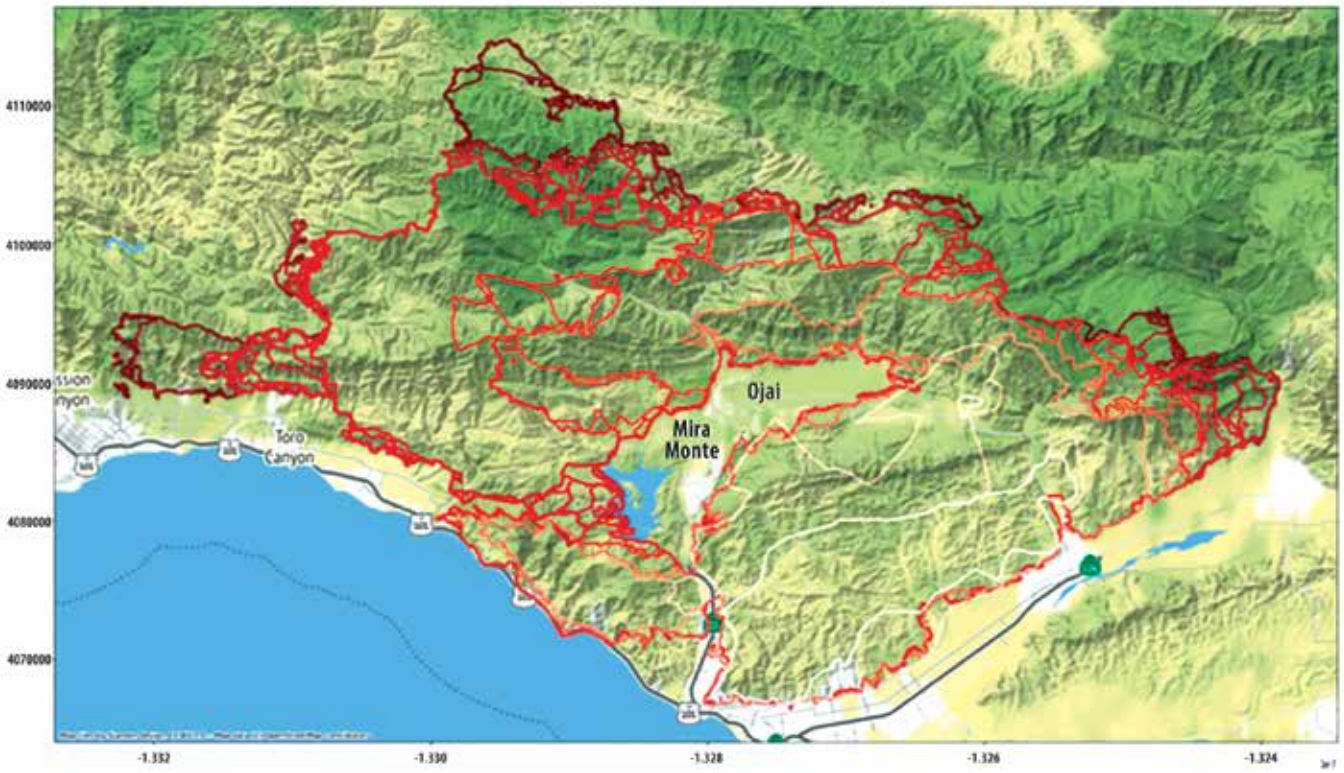
We know data is and has been used to obfuscate and distort the realities of these communities. Data aggregation and summarization represent individuals' lives. We therefore need community members to know their stories are embedded in these data findings and that exclusion of this interpretation is an ethical issue. Community members need to understand how to investigate why data is absent because, in so doing, they can understand whether exclusion occurred by chance or through systematic practices and how this may potentially bias estimates.

For this reason, the data and research literacy courses pay keen attention to research and data ethics. We emphasize the importance of being ethical in the stories you choose to tell and how you choose to tell them. Ethical issues around community data as it relates to ownership, sharing, privacy, identification, and commercialization have real-life implications.

New forms and types of data collection have created many ethical questions that remain unaddressed. The American Statistical Association supports the ethical use of statistics and data science to ensure policies and ideals do not harm, marginalize, and otherwise divide people and groups within our society. ■

Statisticians and Wildfires at the Wildland/Urban Interface

Christopher K. Wikle and Jonathan R. Bradley



Fire perimeter evolution for the Thomas Fire from December 2017–January 2018 in southern California; lighter lines correspond to earlier times.

Millions of acres of land are destroyed by wildfires every year, and they pose a significant threat to humans in terms of property damage and potential loss of life. Of course, there is a significant effect on the ecology of areas affected by wildfires, as well.

The risk of devastating fires and their costs have increased due to trends in land development near the boundary between wildland and urban areas, which is known as the wildland urban interface (WUI). Additionally, there is evidence that wildfires are becoming bigger and more common, and this trend is predicted to continue because of global warming.

This is presenting ever greater challenges to those who must manage, mitigate, and predict fires at the WUI. In addition to the increasing economic burden, the main responsibility during an active wildfire is to protect or evacuate people and mitigate the effects on property and infrastructure. Therefore, it is crucial to develop trustworthy models that can anticipate potential dangers from imminent wildfires to human populations and offer a mechanism for quantifying uncertainty that enables the creation and evaluation of feasible mitigation and prediction strategies.

Although statisticians have been involved in various aspects of wildfire modeling for years, greater numbers of megafires, new data sources, and methodological advancements in statistics and data science are leading to increasing involvement.

Wildfire risk assessment and modeling at the WUI provides a perfect example of the need for engagement from a variety of stakeholders, including federal, state, and local governments and NGOs. Indeed, these various stakeholders often have competing interests and objectives, which leads to substantial disagreement about managing forests and mitigating fire risk.

One consequence of this uncertainty is the development of more systematic and empirical-based

Christopher K. Wikle is Curators' Distinguished Professor of Statistics at the University of Missouri with additional appointments in soil, environmental, and atmospheric sciences and the Truman School of Public Affairs. His primary research interests are in spatiotemporal statistics applied to environmental and ecological processes, with particular interest in dynamics. He is a fellow of the American Statistical Association, Institute of Mathematical Statistics, and International Statistical Institute and has received the ASA Statistics and the Environment Section Distinguished Achievement Award.

Jonathan R. Bradley is an associate professor in the Florida State University Department of Statistics. His primary interests include Bayesian analysis and spatiotemporal statistics with applications to environmental processes and official statistics.

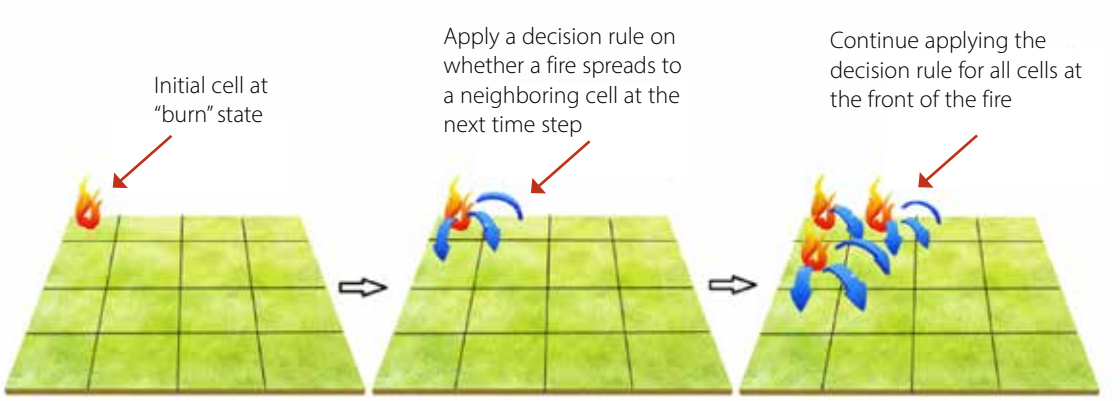


Illustration of a probabilistic cellular automata model for the evolution of a fire on a discrete grid.

. . . there is a greater-than-ever need to understand, predict, and develop mitigation strategies for wildland fires.

assessments of forest resiliency (i.e., the capacity of an ecosystem to rebound or adapt after disturbance). It is becoming clear that this requires an understanding of multiple mechanisms across a wide range of spatial and temporal environmental scales, as well as trophic levels in the ecosystem.

Substantive evaluation of these processes must consider a wide range of data and more sophisticated statistical methods. Fortunately, there are increasingly large amounts of publicly available, georeferenced environmental and ecological data and administrative records available to assist modelers and data analysts in these endeavors.

In addition to forest resiliency, there is still a great need for statisticians to contribute to several fire-specific areas of inquiry. These include predicting fire occurrence risk (e.g., the probability of ignition), fire growth (i.e., rate of spread, size), and fire frequency (e.g., the annual area burned in a region, the interval between fire arrivals on the landscape and burn probability after ignition).

One example of how current advances in statistical methodology are already affecting WUI research is spatial and spatiotemporal extreme value analysis. This is an area, mainly coming out of the environmental statistics community, that is rapidly advancing due to the need to provide greater understanding of the potential for extreme outcomes associated with global and regional climate change. Although the evaluation of wildfire risk has been considered by statisticians for quite some time, the advancements in spatial extremes modeling are providing ever greater understanding of these risks across the landscape. These approaches are now being combined with machine learning models to enhance risk assessment. The topic is of such importance that a wildfire risk data set was featured at the 2021 Extreme Value Analysis conference data competition.

Another major area in wildfire modeling where statisticians are starting to play a bigger role is wildfire spread. Modeling wildfire spread requires an understanding of the pathways in which fires propagate. Generally speaking, there are three ways for a fire to spread: convective heat transfer (where the flame directly contacts a source of suitable fuel); radiant exposure (where heat from nearby large flames ignites suitable fuel); and firebrand shower (or "spotting," when new ignition occurs far from the current fire). The first two mechanisms spread the fire in a somewhat continuous manner, but firebrand shower spotting leads to new fires away from the primary fire front and has been demonstrated to be the primary factor in significant conflagrations at the WUI that causes rapid spread.

The majority of operational wildfire spread models are empirically based, use only a small amount of observational data from past fires, and frequently assume the environment is homogenous and well-known. When flames are burning over

homogenous fuel sources with relatively steady winds and level terrain, such models function reasonably well. They do not, however, work well for intense fires under varying settings (e.g., steep terrain, rapidly changing atmospheric conditions, multiple fuel types, extremely dry conditions) or when the fire itself modifies the local weather.

In these situations, models that combine an atmospheric model with a fire-spread component have shown to be substantially more accurate. Unfortunately, these models are expensive and not currently feasible to run in real-time, do not provide uncertainty quantification, and do not adequately account for spotting.

A longstanding approach to parameterize fire fronts for wildfire spread is representing the front as a level set and advecting it according to some empirical-based rate-of-spread formula. These models have recently been cast in a Bayesian state-space setting and used to assimilate data in real time, as well as to include rate of spread parameter estimation and uncertainty quantification. They can also be placed in a Bayesian hierarchical model framework and, with speed of propagation, considered as a Gaussian process. This allows one to predict the propagation of the fire front and characterize its uncertainty.

Another longstanding approach for modeling wildfire spread is based on cellular automata (CA). CA models dynamically evolve an agent, or automata (i.e., a location referred to as a cell), based on the state (e.g., burning, not burning) of the cell's "neighbors." Although these models are used by many subject-matter scientists, the statistical theory of such agent-based models is an important area that needs development where the vast majority of such models have deterministic transition rules based on empirical analysis. These models have been formulated as stochastic models using a rigorous Bayesian hierarchical modeling framework. However, there have only been a few recent examples of fully Bayesian models in the context of wildfire spread prediction. Hence, developing CA models for wildfire spread under a formal Bayesian uncertainty quantification (UQ) framework is an important emerging area.

Recently, CA models have been combined with deep learning methodology without using formal UQ. The development of CA models for wildfires using deep learning methodology is in line with much of the current methodological development in the prediction of wildfires. Recently, deep learning methodology has been effectively used for predicting the spread of wildfires, predicting susceptibility of fires, categorizing fuel types, and developing decision support systems.

Further Reading

Banks, D. L., and M. B. 2021. Statistical challenges in agent-based modeling. *The American Statistician,* 75(3):235-242.

Cisneros, D., Y. Gong, R. Yadav, A. Hazra, and R. Huser. 2021. A combined statistical and machine learning approach for spatial prediction of extreme wildfire frequencies and sizes. arXiv preprint *arXiv:2112.14920*.

Dabrowski, J. J., C. Huston, J. Hilton, S. Mangeon, and P. Kuhnert. 2022. Towards data assimilation in level-set wildfire models using Bayesian filtering. arXiv preprint *arXiv:2206.08501*.

Falk, D. A., P. J. van Mantgem, J. E. Keeley, R. M. Gregg, C. H. Guiterman, A.J. Tepley, ... and L. A. Marshall. 2022. Mechanisms of forest resilience. *Forest Ecology and Management,* 512:120–129.

Hazra, A., B. J. Reich, B. A. Shaby, and A. M. Staicu. 2018. A semiparametric spatiotemporal Bayesian model for the bulk and extremes of the Fosberg Fire Weather Index. arXiv preprint *arXiv:1812.11699.*

Koh, J., F. Pimont, J. L. Dupuy, and T. Opitz. 2021. Spatiotemporal wildfire modeling through point processes with moderate and extreme marks. arXiv preprint *arXiv:2105.08004*.

Li, X., M. Zhang, S. Zhang, J. Liu, S. Sun, T. Hu, and L. Sun. 2022. Simulating forest fire spread with cellular automation driven by a LSTM-based speed model. *Fire,* 5(1):13.

Taylor, S. W., D. G. Woolford, C. B. Dean, and D. L. Martell. 2013. Wildfire prediction to inform fire management: Statistical science challenges. *Statistical Science,* 28(4):586–615.

Xi, D. D., S. W. Taylor, D. G. Woolford, and C. B. Dean. 2019. Statistical models of key components of wildfire risk. *Annual Review of Statistics and Its Application,* 6(1):197–222.

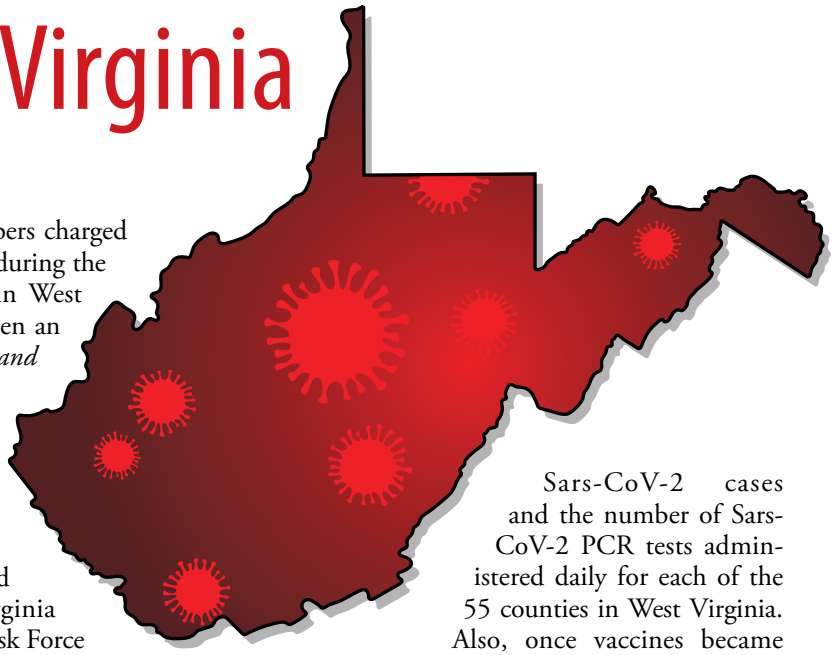
Zhang, L., B. A. Shaby, and J. L. Wadsworth. 2021. Hierarchical transformed scale mixtures for flexible modeling of spatial extremes on datasets with many locations. *Journal of the American Statistical Association,* 1–13.

Given the changing environment due to climate change, ever-increasing human habitation at the WUI, and associated destruction of life and property that is occurring, there is a greater-than-ever need to understand, predict, and develop mitigation strategies for wildland fires. This will require the involvement of governmental and NGO stakeholders, as well as academics. Statisticians are in a unique position to contribute, given the need to account for a vast array of uncertainties and ever-increasing data volumes. ■

Using Data Science in the COVID-19 Pandemic in West Virginia

Brad Price

Brad Price is an associate professor in the management information systems department at West Virginia University's John Chambers College of Business and Economics; co-director of the biostatistics, epidemiology, and research design core and West Virginia Clinical and Translational Science Institute; and chief data scientist for the governor of West Virginia's Joint Interagency Task Force.



As one of the team members charged with working with data during the COVID-19 pandemic in West Virginia, I noticed there has been an overall theme: *Communicate and share the necessary information to keep the 1.79 million residents, health care facilities, and economy of West Virginia protected.*

To do this, groups from all over the state—including the Department of Health and Human Resources, West Virginia Governor's Joint Interagency Task Force on COVID-19, West Virginia National Guard, West Virginia Clinical and Translational Science Institute, and Data Driven West Virginia—partnered to share data and find solutions for West Virginians. The goal of every analysis was to provide pandemic leadership and policymakers with *actionable insights* based on relevant data.

While this may seem obvious, the ability to operationalize resources and the pandemic response because of insights or recommendations is key to gaining the trust of leadership and the public while working in dynamic situations.

Many methods and techniques have been used to deploy testing resources, provide vaccinations, and track vaccination metrics. We continue to develop new methods to adjust for various changes in the disease, policies, and public sentiment. Though these new developments are important, the most critical tool we have is the ability to collect and maintain visibility into the appropriate data for decision-making. Without the appropriate data, the software adage of "garbage in garbage out" will hold.

Many of the analyses performed throughout the pandemic required the number of daily confirmed Sars-CoV-2 cases and the number of Sars-CoV-2 PCR tests administered daily for each of the 55 counties in West Virginia. Also, once vaccines became available, vaccination data for each county stratified by target age groups was available. While information such as this could produce localized public health measures such as the real-time reproduction number (Rt), positivity, testing, and vaccination rates, it could not provide visibility into health care resources in this community.

To do this, the West Virginia Hospital Association provided a daily survey to understand how many patients in each health care facility around the state were COVID patients and how many were being treated with COVID protocols prior to being a confirmed COVID-positive patient. Furthermore, this survey collected information about the number of patients in intensive care units and on ventilators.

During the early phases of the pandemic, when concerns centered on personal protective equipment for frontline workers, a second survey by the West Virginia Hospital Association was built to include questions about amounts of PPE and expected PPE use. Collectively, this data painted an operational picture of resources used for COVID in West Virginia hospitals.

As the pandemic continued, the importance of data sharing became more apparent, especially as West Virginia's leadership made the critical decision to forgo the federal pharmacy program and instead use local pharmacies to prioritize nursing homes and the 60+ population for vaccination. The Joint Interagency Taskforce developed a novel vaccination inventory management system based in Shiny to manage supply and demand to distribute vaccines. This system allowed the multiple entities in the taskforce to share requests for vaccines and see the allocation being distributed around the state in real time. As supply exceeded demand, the system allowed vaccine providers to directly request vaccines and monitor orders and scheduling of vaccinations in their communities.

It became critical to understand the number of vaccinations administered to West Virginians once vaccines were distributed. Data from West Virginia was combined with information from the Centers for Disease Control and Prevention to better understand the demographics of West Virginians who were vaccinated.

The last point we must make about data quality before we can get to the statistical analyses, models, or machine learning aspects of the tools we used is that the COVID-19 pandemic has been dynamic, to say the least. Not only have protocols and behavior changed, but the virus has evolved. The variants present challenges through changes in transmissibility, as well as reacting differently to treatments.

We have seen the virus be dynamic, but the data and policies around how data is collected have also become dynamic. We have moved from mandated testing to facing a negative public sentiment around testing, affecting our ability to understand case counts and positivity rates relative to other times in the pandemic.

The pandemic has created a public health data infrastructure investment that should be kept and maintained for all public health crises, because the first step to any critical response is having access to the right information. *If we have learned anything since March 2020, it is that data infrastructure in public health is part of the national critical infrastructure.*

The statistical tools that have been used during the pandemic response have also had to reflect the dynamic nature of the data. Additionally, a balance has had to be struck between what is scientifically interesting about COVID-19 and what is operationally actionable for decision-makers. While the former may produce publications, the latter provides the necessary analyses to deploy resources that directly affect the pandemic.

For instance, forecasting the number of COVID-19 cases is important to understand, but the question is what resources it affects. Do county-level case counts help deploy PPE to hospitals? Do these case count forecasts provide insights about the number of tests that need deployed to an area? Or do they provide guidance on the number of hospital beds or ventilators that must be available to maintain care for that population? This creates the distinction between the scientific question and the operational question.

We must also make sure behaviors and changes of behaviors throughout the pandemic are taken into consideration. For instance, how people seek care is a key component of describing and preparing for any hospitalization surges that will occur. In West Virginia, we found the way individuals seek care for COVID is similar to how they seek care for emergency room visits, thus we use that as a proxy for how each individual in the state will seek care during any type of surge.

In the early stages of the pandemic, this behavioral type analysis set the base for the suspectable population for the adjusted compartmentalized models we used to develop PPE forecasts. As more data became available and supply chains became more stable, these methods gave way to more traditional statistical and inventory control methods.

Most recently, we have been developing machine learning methods to inform targeted testing events. This tool has been used to deploy testing resources provided by the National Institutes of Health RADX-UP project to localities in West Virginia through community lead events.

COVID-19 continues to affect our communities. As data has become more abundant and the disease has become more dynamic, data analysis has become more difficult. One of the key components to any success thus far has been insights delivered from data. It is of the utmost importance that we continue to develop these methods and the infrastructure required to respond to health crises. ■

Prescribing Privacy: Human and Computational Resource Limitations

What Statisticians and Data Scientists Can Do

Jingchen (Monika) Hu and Claire McKay Bowen

Jingchen (Monika) Hu is an associate professor of statistics at Vassar College. Her research focuses on statistical data privacy methods, mainly synthetic data and differential privacy. She teaches a senior seminar on statistical data privacy and engages undergraduate students in learning cutting-edge methods.

Claire McKay Bowen is a principal research associate in the Center on Labor, Human Services, and Population and leads the Statistical Methods Group at the Urban Institute. Her research focuses on developing and assessing the quality of differentially private data synthesis methods and science communication.

President Biden's day-one Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government committed federal agencies and White House offices to actively pursuing more equitable engagement and outcomes for people of color and underserved communities. However, many statistical agencies do not collect or release detailed demographic data and statistics due to growing concerns about disclosure risks. For instance, people of color with low incomes are more susceptible to privacy attacks because they more heavily rely on smartphones for internet access and provide more personal information for free cell phone app services, according to Mary Madden in the report *Privacy, Security, and Digital Inequality.* Such information collection makes them more easily identifiable, especially if they are located in rural geographic locations.

To address these issues, some public policymakers propose agencies review and update their privacy protection methodologies with more modern data privacy and confidentiality techniques. Yet, many agencies cannot update their privacy protection policies due to the lack of both human and computational resources. This leaves some asking, "Why?" and, more importantly, "What can we do?"

Human Resource

On the human resource side, there exists a gap between the growing demand for professionals in modern data privacy and confidentiality techniques and the training of such professionals at educational institutions and in the workplace. These professionals should be experts in privacy and confidentiality techniques who can design and implement tailored approaches to specific data sets, evaluate the effectiveness of the approaches, and potentially provide training on the methods to colleagues.

The workforce demand spreads across statistical agencies, local and state government entities, and private sector organizations. At the federal statistical agency level, there are trained experts who routinely design, implement, and evaluate the techniques and approaches and sometimes a disclosure review board to make final decisions. However, in many other agencies with fewer resources, such as local governments, the data privacy and confidentiality work requires establishing consulting relationships with privacy and confidentiality experts and a disclosure review board is far away from being created. At private sector organizations, large and small, active recruiting of trained experts in data privacy and confidentiality has been ongoing despite stalled recruitment efforts overall.

A search on LinkedIn with the keyword "privacy" showed more than 210,000 results at the time of this writing, which includes privacy, security, and decentralized learning at Microsoft Research; privacy engineering at Amazon Business; information governance and privacy at PwC; and privacy solutions architect at Google. This search alone demonstrates the enormous demand for all types of data privacy experts, such as those in privacy law, cybersecurity, and statistical privacy.

When it comes to training professionals in data privacy and confidentiality techniques, little is happening in statistics and data science, especially at the non-PhD levels. Most of the data privacy and confidentiality courses focus on differential privacy and appear in computer science PhD programs for graduate computer science students. At the undergraduate level, there are occasional seminar courses taught by professors who conduct research in the area.

These advanced-level courses would cover the nuts and bolts of learning and implementing the techniques, although not necessarily the theoretical underpinnings. Yet, given the growing interest and demand, most undergraduate courses on data privacy and confidentiality are at the introductory level, not necessarily designed and taught by professors trained in this area and open to students from all backgrounds.

This means students in these courses typically do not get into the details of how to perform the techniques in practice. Nevertheless, it is encouraging to see that many technical online courses on this topic are being offered for professionals, which again demonstrates the enormous workforce demand to train more professionals.

Computational Resource

On the computational side, not having readily available computational tools will hinder the accessibility for professionals to implement more modern data privacy and confidentiality methods. They might not have the proper computing environment to run these methods or the technical background (expert knowledge and/or programming skills) to hand code them. Moreover, hand coding is more prone to errors and might be less efficient.

As mentioned before, trained statistics and data science professionals in data privacy and confidentiality should understand the nuts and bolts of these methods, but not necessarily the theoretical underpinnings. For example, we do not need to know how to build a bike in order to ride it.

If all we need are bikes, then are there enough bikes for people to ride? Unfortunately, few bikes exist.

In the statistical field, one software tool is synthpop, an R package that implements synthetic data generation and creates a 'fake' data set based on a statistical model that aims to have the same statistical features and data structure as the confidential data. The synthpop R package also measures data usefulness. However, it lacks the functionality to evaluate the level of protection the generated synthetic data sets provide.

Another research group out of Harvard University started OpenDP, which it describes as "a community effort to build trustworthy, open-source software tools for statistical analysis of sensitive private data." OpenDP has partnered with Microsoft, engaged with the broader data privacy and confidentiality community, and created a GitHub repo. However, their platform is still under development and not ready for primetime.

Despite the demand for data privacy and confidentiality software, challenges in developing these tools include not enough funding and time to support this type of work.

Dig Deeper

Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government *https://bit.ly/3QrJHJU*

Privacy, Security, and Digital Inequality *https://bit.ly/3JLdjzN*

Differential Privacy: What Is It? *https://bit.ly/3dknbED*

OpenDP
https://opendp.org
https://github.com/opendp/opendp

synthpop
https://bit.ly/3SypG6x

What can we, as statisticians and data scientists, do to address these resource challenges?

There is much that could be done to advance the field. Following are a few we recommend starting with:

- Incorporate data privacy and confidentiality into undergraduate curricula that goes beyond the basic introduction, such as applying appropriate methods to real data and evaluating their effectiveness.
- Take more of a presence in the space through research, teaching, and science communication. It often feels like 1 to 20 for statistics vs. computer science.
- Focus on how to translate theory to applications and deployment, rather than only the theory.
- Advocate for more funding for applied research and deployment (i.e., computational tools and educational resources), instead of only on new method development.

We think taking these steps alone will not solve all the human and computational resource limitation problems, but they will help alleviate them. ■

SUPPORTING COMMUNITY-ENGAGED RESEARCH

Ethical Challenges and Plausible Responses in Statistics, Data Science Practice

Rochelle Tractenberg

Rochelle Tractenberg is a tenured professor in the Georgetown University departments of neurology; rehabilitation medicine; and biostatistics, bioinformatics, and biomathematics. She is an ASA Accredited Professional Statistician and a fellow of the ASA and AAAS. She earned a PhD in cognitive sciences from the University of California, Irvine and a PhD in measurement, statistics, and evaluation and doctoral-level certificate in gerontology from the University of Maryland, College Park. She chaired the ASA Committee on Professional Ethics from 2017–2020 and the working groups on revising the ASA Ethical Guidelines for Statistical Practice in 2016 and 2018 (co-chair in 2021). She is the author of two forthcoming books about ethical reasoning and its application with the ASA ethical guidelines.

The American Statistical Association first published its Ethical Guidelines for Statistical Practice in 1995. Since 2016, when the first formal revision effort was approved by the ASA Board, they have been slated for review and revision every five years.

While the ASA Committee on Professional Ethics maintains and disseminates the guidelines, their actual application has been a challenge for the committee and entire ASA community.

Critically, the guidelines are described as promoting *ethical decision-making.* The guidelines specify that "statistical practice" includes activities such as designing the collection of, summarizing, processing, analyzing, interpreting, or presenting data, as well as model or algorithm development and deployment. They also explicitly point out that, irrespective of job title, level, or field of degree, the guidelines apply whenever the individual engages in statistical practice.

Ethical reasoning is a process that can be learned and improved. The six knowledge, skills, and abilities of ethical reasoning based on a mastery rubric by Kevin FitzGerald and me are the following:

1. Determining your prerequisite knowledge
2. Identifying decision-making frameworks
3. **Recognizing an ethical issue**
4. **Identifying and evaluating alternative actions**
5. Making and justifying decisions
6. Reflecting on the decision

With a bit of luck, the ethical statistical practitioner will need only #1 and #2 for 95 percent of their work. How to use the ASA Ethical Guidelines to practice ethically is the topic of most of my book *Ethical Reasoning for a Data Centered World.* Here, we treat the ASA Ethical Guidelines as the "prerequisite knowledge" needed to identify when ethical challenges arise. Memorizing the guidelines is not the focus; knowing their role and the general organization is enough to start with.

The ASA Ethical Guidelines for Statistical Practice includes the following eight core principles and an appendix for organizations and institutions with 72 specific elements:

- Professional integrity and accountability (12)
- Integrity of data and methods (7)
- Responsibilities to stakeholders (8)
- Responsibilities to research subjects, data subjects, or those directly affected by statistical practices (11)
- Responsibilities to members of multidisciplinary teams (4)
- Responsibilities to fellow statistical practitioners and the profession (5)
- Responsibilities of leaders, supervisors, and mentors in statistical practice (5)
- Responsibilities regarding potential misconduct (8)
- Appendix: Responsibilities of organizations/institutions (12)

In terms of ethical decision-making frameworks (#2 on the list), there are two that can be helpful in accomplishing #3, recognizing an ethical issue. These are "utilitarian" and "virtue" perspectives. The virtue ethics perspective can generally be summarized as, "what would the (ideal) ethical practitioner do in this situation?" Since the ASA Ethical Guidelines elements have the stem, "the ethical statistical practitioner (does x)," it is clear the guidelines feature the virtue perspective.

However, the utilitarian perspective can generally be summarized as, "how can benefits be maximized while harms are minimized in this situation?" The guidelines can also be used to check whether there may be harms when responsibilities outlined in the guidelines are ignored.

Thus, we see that not only are the ASA Ethical Guidelines part of prerequisite knowledge, but the actual event or behavior also must be clearly described.

Many ethics cases are purposefully vague to help encourage discussion. However, to determine what is going on and how best to respond when unethical behaviors occur or are observed, it is important to be able to clearly describe exactly what is going on/happened.

The most straightforward ways to identify an ethical challenge in any given situation are the following:

1. Determine if, once the actual event or behavior is clearly described, it is inconsistent with one or more of the ASA Ethical Guidelines principles or elements.
    - If one or more principle or element is violated (or threatened), there is a problem and you've moved closer to identifying how to address it.
2. List the harms—both actual and potential—as well as any benefits of the event or behavior for all stakeholders.
    - If there are more harms (actual or potential) than benefits, harms are more serious than benefits, or harms accrue to vulnerable or at-risk populations (especially if benefits do not accrue to these populations), then you've identified the problem. The stakeholder analysis does not move you closer to identifying how to address it, but you will have evidence that your solution was successful if the harms you identified are mitigated or eliminated.

A stakeholder is defined as "one who is involved in or affected by a course of action" by Merriam-Webster. In the context of ethical case analysis, the stakeholder is simply any individual or group that might be affected by the outcome of the event or behavior. Stakeholders include yourself (and your professional reputation), your boss, your employer/organization, your nonstatistical practitioner colleagues, the profession (of statistics/data science), and the public. Note that you can use a stakeholder analysis to evaluate your alternative actions and determine if an event or behavior may constitute an ethical challenge.

At this point, we can see the guidelines are important for ethical reasoning steps #1 and #3; they can be important in #2 (to support a virtue ethics approach), but can also be helpful in identifying *stakeholders* and harms/benefits that accrue to each. Note that Principle C outlines responsibilities to stakeholders, while principles A (*yo*u), D ("anyone directly affected by statistical practices"), E (team members), and F (other practitioners and the profession) are each explicit about responsibilities to different stakeholders.

Once an ethical challenge has been identified, there are **always three** alternative actions to take or decisions that can be made:

- Do nothing (ignore the unethical behavior), ignore the request to do (the unethical thing), or agree to do (what was asked, even if that conflicts with the guidelines or creates more harms than benefits for any stakeholder).
- Consult with a peer or supervisor using the professional guidelines or other resources.
- Refuse to do what was asked (if it was unethical) and/or report violations of policy, procedure, ethical guidelines, or law.

The ASA Ethical Guidelines exist to help practitioners identify and avoid—or prevent/put a stop to—unethical statistical practices. Many elements of the guidelines explicitly state the ethical practitioner avoids, avoids condoning, and avoids *appearing to condone* unethical behavior. Thus, it is never ethical to "do nothing" when faced with an ethical challenge. That is a decision and a "plausible alternative," but it is never ethical.

To consult a peer or supervisor will be much more straightforward if you have completed the ethical reasoning steps 1–3 as outlined here. That is, when you determine who would be a good person to consult with, you simply use your reasoning steps 1-3: "Hi, colleague, I'm in a bit of a situation. I was directed to do/observed X, and that's contrary to the Ethical Guidelines [list]. What do you reckon I should do?"

Moreover, if you have identified exactly what the ethical challenge is and what elements or principles in the guidelines are—or would be—violated by the behavior you observed or were asked to do, your and your conferee's decisions about what to do *will be clearer*. For example, "stop <doing what is contrary to the guidelines> and do <what the guidelines specify the ethical practitioner does instead>." And "share the guidelines with <bad actor #6> so they stop asking the statistics practitioners to do <guideline violations a, b, c …>."

Finally, making a decision (Step #5) requires both the decision *and* its justification. This is clearly supported by the effort you put into steps 1–3, and since you'll never choose alternative Option a (do nothing/ignore/fulfill the unethical request), there are two clear options to tailor to your situation (Option b: confer; Option c: report).

Ethical reasoning and the 2022 ASA Ethical Guidelines are specifically formulated for *all* practitioners at all levels. Ethical practice standards are not "moral principles," but specific descriptions of what constitutes ethical practice. The implication is that if a practitioner does not follow the guidelines/code, then they are not doing their job ethically.

The guidelines support professional and scholarly and scientific work in, and with, statistics and data science. When practitioners do not follow ethical practice standards, all those who make decisions on the basis of the results of quantitative practice may find their decisions or scientific or scholarly work undermined. ■

A Date with Data: Stepping Toward Data Literacy

Nairanjana (Jan) Dasgupta

Nairanjana (Jan) Dasgupta is a Boeing Distinguished Professor at Washington State University, a fellow of the American Statistical Association, and chair of JEDI WNAR International Biometric Society. This year, she was named Woman of the Year at Washington State University for her tireless work in advancing statistics and increasing opportunities for women in the field.

Data is considered the new "liquid gold." And phrases like "deep dive into data," "data culture," "data-based decision-making," and "data-informed decision-making" have become ubiquitous among decision-makers. With the rise of these phrases, another phrase is also becoming fashionable: data literacy.

The dictionary says "data" = information and "literacy" = ability to read and write. So, in the most literal sense, the phrase could mean the ability to read and write with data.

There are gaps in access to training in data science, with unequal access across racial and ethnic groups. The issue of unequal access exists in all STEM fields, but it is exacerbated in data science, as unequal racial representation translates to algorithm writers who are nondiverse and unrepresented communities that are victims of algorithmic bias.

There are no easy solutions. The ultimate answer is increasing the pipeline by focusing on exposure to and interest in data from an early age. We can do this by introducing data concepts in elementary school and keeping them separate from mathematics.

Math is generally arithmetic in the early grades. Data literacy is more about number sense, storytelling, and visualization. A kindergartener may not be able/want to add 20 numbers in a minute (how math is often taught in underprivileged schools) but their favorite color of lollipop is something they can understand and relate to. They can understand that bringing more red lollipops to school (if that is the most common favorite color) is a data-based decision.

Teaching data as a concept outside math is a tough sell for math educators. There has been much research on incorporating data into math, but unless there is buy-in from teachers (especially in under-funded schools), it will be accessible to only a few groups. Math education is an established field; data education is in a fledgling state.

To advocate for data literacy, I spoke with math coordinators across various education districts in Washington (a tech-friendly state). The most common response I received was that data is already part of the curriculum and teachers are

Seven-year-old George, who wants to design computer games, puts his family poster together with his family, friends, and facilitator Kirsie.

overstretched and underpaid. Essentially, "Thanks but no thanks. We do not have the funding or bandwidth to do what you are asking."

Another oft-repeated response was that teachers run out of time before they get to the data chapter because they have to prepare students for testing.

As an alternative, I came up with a data literacy initiative. I would partner with community organizations, go to underserved communities, and lead "Date with Data" evenings. I wanted to target elementary children but invite the *whole family* in the hope of having a conversation about data.

The idea was to have an evening of fun and games while using the word data in all the activities. We served food, as it took away the burden of cooking dinner for the families. Having the entire family allows the parents to have child care for all their children. The parents and older siblings are crucial, as we want to talk to them about how data can be used and misused. The whole idea is to spike an interest in data, start a conversation about it, and create awareness within families and the community.

A Date with Data: Cashmere and Tri Cities

Both camps were held from 4:30 p.m. to 8:30 p.m. (after work for these communities) on May 19 and 21 in Cashmere (a large migrant community) and Richland (a large Latin community). The community partners were the Cashmere School District Federal Programs and the Hispanic Chamber of Commerce of Tri Cities. The second camp was funded by Amazon.

The program directors recruited families with elementary children. To act as facilitators for the camp, I recruited seven graduate students, three undergraduate students, and a high-school student with backgrounds in statistics, data analytics, comparative languages (fluent in Spanish), and psychology.

We had 35 attendees in Cashmere (11 children, 24 adults) and 23 attendees in Richland (13 children, 10 adults). The Cashmere event was in the elementary school cafeteria, and the Richland event took place at the WSU Tri Cities campus building. The attendance was lower in Richland because the building was on campus and a bit hard to find. It was also raining heavily that day.

Facilitator Feedback

What did you really enjoy?

ND: I enjoyed the M&M guessing game.

SM: Interacting with parents to learn about the community and its families.

DR: I enjoyed the family data poster presentation.

JN: I really had a good time with the two girls, knowing how smart they are.

PB: I really enjoyed playing games with the kids and talking to them about their futures.

JD: The car races with the kids.

KL: Watching the kids present the posters.

SM: All the kids' impressive artwork. Angeles and Jinesh's children's interest in the data scrape activity.

DR: Talking about the birthday histogram with the families.

SC: I enjoyed the birthday game.

What was your high point?

ND: Jaime emailing me saying he wants to go to college.

KL: Helping George open up and engage with the activities.

SM: A 5-year-old participant (who could barely write his own name) writing out *"los datos"* (adorably misspelled) great big on his family's poster board.

DR: Also helping George open up with the data scraping activity.

SC: While playing UNO, Josh said he will win and soon after he asked, "What is probability? How do you define that?"

Facilitators: David Rice, Justice Nii-Aytey, Shivani Sawant, Swarnita Chakraborty, Meghna Johnson, Kirsie Ludholm, Payton Bokowy, Molly Behrends, Freeman Chen, and Spencer Martin

At each table, we had similarly aged children with at least one facilitator and a mason jar filled with M&M candy. The icebreaker was guessing the number of M&Ms in the jar. Whoever guessed closest won the jar and M&Ms, but all the kids got some, as did the parents. They then filled out the following questionnaire:

My birthday is on _____

My favorite color is _____

My favorite animal is _____

My favorite time of the day is _____

My favorite song is _____

My favorite season is _____

My favorite food is _____

My favorite number is _____

My favorite candy is _____

My favorite ice cream is _____

We pre-filled posters with the months and birthdays and then constructed a histogram. We asked the kids which month they thought most birthdays were in. We talked about the most common birthday and conjectured about why it was.

Next, we had the kids design a game of putting table tennis balls into jars and coming up with a point system. We also had toy car races. The kids each picked a car and parents, kids, and facilitators predicted outcomes. We tried to reason why one toy car would be faster than another.

Josh and Divya make posters with their parents and facilitator Shivani.

We played UNO and dice games and asked strategy questions. We emphasized to the kids that they were making "data-informed" decisions in their plays.

We plotted the sum of two dice using stickers. We talked about what was likely and what was not.

Dinner was a big attraction. In each case, we served a taco feast from a local business. In Cashmere, there were more adults than kids. Each facilitator talked with the parents and children during the meal about their future plans and why it was important to understand data in this world of automation.

After dinner, each family sat together and used the information we had collected to visualize their answers however they wanted. Most chose to focus on their favorite animals and drew them using their favorite colors. Parents and kids joined this activity with energy. Then, one kid from each family was invited to talk about their family's "data story." Most of the time, all the siblings came up to present, collectively holding the poster. Often, the parents came, as well. The parents seemed to have more fun than the kids, critiquing each other's art.

The last activity was a "web scraping" from Twitter that we used to create word clouds. The kids and parents typed in their own search words. The kids searched words like "soccer," "football," and "Nemo," while the parents searched words like "jobs," "salary," "Cashmere," and "Tri Cities." It was eye opening to the families to see how much can be done and how easily.

In the end, each family was given a goody bag and left smiling. Many asked us to come again in the fall. Alex, 10, told us he would "come to WSU and study data analytics" and drew the university logo on his poster. He said he would be the class of 2032. Then, Jaime, 14, said, "I want to go to college, but how do I go?"

Participants in both camps were asked if they wanted to do this again and 100 percent said yes. They also said they would bring more of their friends and family next time. We have two more camps planned this summer and, with funding, we should be able to do more in the fall.

The objective was to spark interest and engage in hands-on data literacy. The data from the pilot camps showed we accomplished that. I hope we can continue and really make data the fourth dimension of literacy. ■

Equity and Bias in Algorithms: A Discussion of the Landscape and Techniques for Practitioners

Emily Hadley

Emily Hadley is a research data scientist at RTI International. She is experienced in machine learning, predictive modeling, and natural language processing and is passionate about ethics, bias, and antiracism in statistics and data science.

With the growing use of algorithms in many domains, considerations of algorithmic bias and equity have far-reaching implications for society. Algorithmic bias emerges from the concern that algorithms are not simply neutral transformers of data but compounders of existing societal inequities, particularly when performance is substantially better or worse across mutually exclusive groups.

Algorithmic bias can occur as a result of decisions made throughout the algorithm development and deployment process. Left unaddressed, it can deeply affect equity. While the meaning of equity has been contested for millennia, it is generally considered to be a property of fairness of a decision-making agent or institution.

We already live in a world in which biased algorithms are affecting individual livelihoods. These effects have been described in books like *Weapons of Math Destruction,* in movies like *Coded Bias,* and in the growing Artificial Intelligence Incident Database (AAID). The following have occurred as a result of biased algorithms and algorithmic decision-making:

- Millions of black hospital patients received lower risk scores than similar white hospital patients, potentially excluding them from extra care
- Older workers never saw some age-targeted job ads on Facebook, denying them potential work opportunities in a direct violation of US employment law
- Faulty facial recognition led to the arrest of a man for a crime he did not commit

Few and narrow laws exist in the US to govern algorithm development and deployment or address algorithmic bias. Anti-discrimination laws cover the use of specific protected classes like race and sex in algorithms in domains such as lending and employment. The Federal Trade Commission recently took a role in holding companies accountable for deceptive and discriminatory algorithms that affect consumers, and individual states like California and New York have developed legislation related to algorithmic data inputs and the use of algorithms. Yet, there remains no comprehensive legal guidance in the US for algorithm development and usage.

Thus, it falls to the practitioners building these algorithms—including statisticians and data scientists—to interrogate decision-making throughout the algorithm development process and identify opportunities to enhance equity.

Techniques for Practitioners

There is no single tool or approach that makes an algorithm unbiased. Rather, practitioners should adopt a commitment to recognizing opportunities for bias in decisions throughout the development process and act to address these challenges when possible. The following are four techniques related to development:

Technique 1: Advocate for Representativeness of Data

Representativeness of data is a topic often covered in an introductory statistics course. Students learn it is often inappropriate to make sweeping generalizations of results from a limited data set and apply them to populations for which that data is not representative. Yet, in practice, representativeness is often not prioritized and leads directly to biased algorithms.

Facial recognition data sets used to build tools for social media companies and law enforcement have historically skewed white and male, leading to less accurate predictions for non-white and

non-male individuals. Health care AI systems are overwhelmingly built using data from just three states (CA, NY, and MA), and this lack of data diversity is likely contributing to biased health algorithms.

When developing an algorithm, practitioners should analyze the representativeness of available data in comparison to the population of interest, identify disparities, and advocate for greater data diversity.

Technique 2: Interrogate Use of Sensitive Attributes in Algorithms

Sensitive attributes are protected characteristics like race, sex, or age for which bias in the algorithm could lead to inequitable decisions. A common argument is that an algorithm must be fair and unbiased if it doesn't include these sensitive attributes, known as "fairness through unawareness." Yet, literature has shown this argument does not hold due to the numerous and often opaque relationships of sensitive attributes with seemingly "neutral" predictors.

The algorithm where black patients received lower risk scores was one that was purported to be fair because race was not included as a predictor; however, it emerged that projected cost—another predictor—was correlated with race, which led to the discriminatory outcome.

Practitioners should recognize it is a myth that simply withholding a sensitive attribute from an algorithm will make it fair. The appropriateness of using a sensitive attribute in an algorithm should depend on the context in which it is used, and, regardless of use, practitioners should evaluate fairness (Technique 3) when possible.

Practitioners can further interrogate their use of sensitive attributes with the following questions:

- How was missingness addressed for the sensitive attributes? Was the approach ethical and thoughtful? How might the approach affect the outcome?
- Was a grouping technique such as combining groups with small numbers used for the sensitive attributes? What assumptions were made in this grouping? Who is prioritized by the grouping?
- Were groups combined to create an "Other" group? This group may not be meaningful for analysis; whose insights will be lost by inclusion in the "Other" group?
- Is a reference level used in the analysis? Why was this reference level selected, and who is prioritized by selecting this reference level?

Technique 3: Evaluate Fairness in Algorithms

Evaluating algorithms for fairness is an evolving area of research. One reason is the competing definitions for fairness. Some metrics align with *individual fairness* such that individuals with similar attributes should have similar outcomes. Other metrics align with *demographic parity* such that the distribution of positive or desirable outcomes should mirror that of the general population class distribution. Still others align with *equal opportunity* such that there should be equal true positive rates across classes.

Selection of fairness metrics should include conversations about the tradeoffs between these definitions with subject-matter experts and those most likely to be affected by the algorithm. Practitioners should prioritize calculation of fairness metrics in their algorithm development workflow.

Technique 4: Consider an Algorithmic Review Board

Academic researchers may be familiar with the institutional review board, an administrative body established to protect the rights and welfare of human subjects. Given the effect of algorithms on individuals, there is increasingly a call for algorithmic review boards at companies developing algorithms with human impact.

Large tech companies and financial institutions have already begun exploring how an ARB or similar committee would work in practice. These company watchdogs can serve as an internal mechanism to evaluate the practices used to collect data and build and deploy the algorithm.

Practitioners should consider if an ARB or similar committee may be appropriate for their own organization.

Statisticians and data scientists are actively involved in developing algorithms being used to make decisions that affect individuals, often with little or no legal oversight. By incorporating individual techniques into their own work, these practitioners can contribute to key decisions that reduce algorithmic bias and improve equity. ■

Q&A

These statisticians and data scientists have been working on the front lines to improve our communities, so we asked them to talk about their current projects; who inspired them; and how to get started supporting where we each live, work, learn, and play.

LEONOR SIERRA has more than 10 years of experience in science communication and policy, with a focus on helping scientists get involved in and affect public debates about science. After earning her PhD in physics at the University of Cambridge, she worked for the UK nonprofit Sense About Science, then as a press officer at the University of Rochester. She is currently a freelancer based in Athens, Georgia. She also collaborates as a Sense About Science associate, most recently on a project about how risk know-how helps communities navigate risk information and assess benefits and trade-offs within their own context.

Who or what inspired you to study statistics/data science?

I haven't studied statistics or data science. My background is in physics, but it's been my work in science communication and risk with different communities that has made me realize just how much of our understanding relies on being able to question statistics.

Describe a current project and tell us what inspires you in your current work?

I have been working on risk know-how (*www.riskknowhow.org*) and am inspired daily by people who take on responsibility for helping their communities understand and engage with different issues.

What is one job skill you learned recently, and why do you find it important?

Recently, I had to take a medical leave. I don't find it easy to take a step back, but sometimes it's important to know when to slow down and recharge.

What are three pieces of advice you would offer your younger self?

- Leaving academia does not mean leaving science.
- You don't have to have decided what you will do in the future; just keep learning.
- Listen and read even more before forming an opinion.

What do you love most about your job?

Talking to people from all over the world, from leading experts to people doing amazing work in their local communities. I learn so much from them and am inspired to do more.

In a President's Corner, Kathy offered this definition of community analytics: bringing the best of statistical science—in collaboration with municipal governments, universities, local businesses, NGOs, and community organizations—to improve lives through a better understanding of our communities and how we live, work, learn, and play. Looking to the future, what do you see as emerging areas and opportunities?

I think there is a huge opportunity to include community practitioners' experience and expertise, especially around risk communication—not just as receivers of information but as experts in their community.

What are the most important skills for students to be developing right now in school to be ready for the future in data science?

I think communication is such an important skill for any scientist. I would encourage them to make the most of any opportunity for workshops, trainings, etc. Also, thinking more broadly about public engagement: how to involve anyone who would be end users of the data or part of the group about which data is being collected and how to do so meaningfully and not as an afterthought.

Name one or two favorite blogs or books you have read and would recommend to others.

The Tiger That Isn't by Michael Blastland was probably the first popular statistics book I read many years ago, and so many of his examples and potential pitfalls have stayed with me.

Superior: The Return of Race Science by Angela Saini is not a book about stats or data science, but I believe it has some important cautionary tales about the data we collect and use—not always judiciously. ■

SUSAN PADDOCK is chief statistician and executive vice president at NORC at the University of Chicago. NORC is an independent research organization that delivers insights and analysis in five principal areas: economics; education; global development; health; and public affairs. NORC statisticians and data scientists work cross-functionally across those areas. Paddock is generally responsible for the methods of design and analysis used in NORC projects and the NORC corporate research and development enterprise. She earned her PhD in statistics from Duke University and her BA in mathematics and biostatistics from the University of Minnesota.

Who or what inspired you to study statistics/data science?

I have always been interested in what happens in our society. In my teens, my interests included politics, human rights, and women's issues. I brought these interests to college and explored majors I thought might lead me to work on those topics. Part of my exploration included taking courses in health policy and women's health, which led me toward public policy.

The backdrop to this was the intense health care reform debate of the early 1990s. One class assignment focused on highly experimental and expensive cancer treatments for patients without good treatment choices. There was information to consider about clinical trials and evidence, health economics, and ethics. The importance of using data to understand such a complex scenario became clear to me. An epidemiologist taught the women's health course. It was my first exposure to identifying gender and racial bias in health research and to scientific practices such as literature reviews and epidemiological study design.

All of this inspired me to seek out professors working in public health, including then-director of undergraduate studies in biostatistics at the University of Minnesota Anne Goldman, who convinced me that majoring in biostatistics and math would prepare me well to bring data to bear on improving health and well-being in our communities.

Describe a current project and tell us what inspires you in your current work?

I am inspired by the challenge of collecting actionable, timely, and high-quality data to guide decision-making and better understand our communities.

My NORC colleagues and I have been developing and applying statistical methods and data science techniques to surveys, administrative records, bibliographic data, environmental sensor data, social media data, etc. Each of these data types has its own strengths and limitations. Recognizing that, we develop approaches to integrate data from multiple sources. This is useful when one data set alone cannot provide valid or precise estimates—such as for obtaining small-area estimates for communities or subpopulations of interest.

At NORC, we conduct surveys that provide data to the public to examine many topics of interest to communities, including COVID-19 vaccination, early child care and education, career trajectories, and voter preferences and attitudes.

One factor that makes all survey data collections challenging is nonresponse. Data science has enabled implementation of approaches to reduce the risk of nonresponse bias through adaptive and responsive sampling designs. This involves using information collected prior to or during data collection to change the design as needed to improve the representativeness of the final sample. This approach can be embedded into a survey data science workflow that, in addition, allows one to ingest, process, analyze, summarize, document, and finalize data sets in a way that is efficient and reproducible.

What is one job skill you learned recently, and why do you find it important?

One of the broad job skills I always aim to improve is communication. In a one-on-one meeting, I work to develop an understanding of my colleague's perspective and how we might communicate best with one another.

There are people who do not like small talk, and then there are others for whom conversations about hobbies, weather, and weekend activities make them feel more connected to the organization. Some people are systematic and careful in their communication approach, while others make exciting and sweeping pitches.

One of the things I do as a leader is to figure out what it will take to create an environment in which people can succeed. Meeting others where they are when we communicate is part of that.

What are three pieces of advice you would offer your younger self?

First, assume less and ask more questions. If you're new to the field, your colleagues are hoping and expecting you'll ask some questions, because we've all been 'new.'

Second, people who are starting out in their careers should pay attention to their full working environment. Not only is it important to expand your skills as a statistician or data scientist, but also appreciating the substantive application area and understanding the mission of your organization will make your career go more smoothly.

Third, get involved in professional activities for career growth, personal development, and networking.

What do you love most about your job?

I love putting data and research findings into the world that people can trust and use to make sound decisions and assessments. Conquering a new statistical problem can be thrilling; there's nothing like putting together the pieces of a puzzle. As a leader, it is rewarding to create an environment that is dynamic and full of opportunities for others to do impactful and fulfilling work.

In a President's Corner, Kathy offered this definition of community analytics: bringing the best of statistical science—in collaboration with municipal governments, universities, local businesses, NGOs, and community organizations—to improve lives through a better understanding of our communities and how we live, work, learn, and play. Looking to the future, what do you see as emerging areas and opportunities?

There are many opportunities! I'll mention just three. First, statisticians and data scientists can empower communities and organizations to make full use of their own data. At NORC, a team developed the Data File Orientation Toolkit (*https://bit.ly/3bHx7re*), which is an open source toolkit for researchers and analysts—particularly those at state and local agencies—to assess the quality of administrative data files for conducting policy analysis.

Second, we have opportunities to make data easier to use by the public. For example, anyone with access to a web browser can use the General Social Survey Data Explorer (*https://gssdataexplorer.norc.org*) to investigate the concerns, experiences, attitudes, and practices of US residents throughout the last 50 years. COVID-19 dashboards empowered many people and communities to monitor data in real time and make personal and policy decisions based on such data.

Third, there is increasing awareness of the importance of stakeholder and community engagement in study design, data collection, and measurement to achieve research and analysis outcomes that are meaningful to communities and promote justice, equity, diversity, and inclusion.

What are the most important skills for students to be developing right now in school to be ready for the future in data science?

I never regretted taking a couple of computer programming courses before I went to graduate school, not only for the obvious benefit of being able to write fast code but also to learn about programming in general. It also helps for students in the job market to have some level of proficiency with one of the major statistical packages and familiarity with others.

Rigorous statistical training remains the compass for statisticians and data scientists to navigate the myriad messy data scenarios out there and—even better—design studies to reduce the messiness as much as possible. Given that messiness, I'd recommend at least some foundational courses related to study design—such as survey sampling, experimental design, or randomized trials—because knowing 'good' study design will help one better cope with problematic scenarios in the future.

Machine learning can be applied to numerous problems, so it is useful to know about it. Knowing how to work with data is important, especially being able to identify anomalies and ask the right questions to understand the quality and meaning of the data. This, of course, means students should have opportunities to work with data sets that can be used to answer questions of scientific interest and, ideally, work as part of a team to gain collaboration experience.

Name one or two favorite blogs or books you have read and would recommend to others.

Twitter (in particular, #statstwitter, #econtwitter, or #rstats) is the fastest way for me to keep up with new developments in the field. I find so many interesting reports, papers, blogs, and 'tweetorials' that way.

On the leadership side, I was excited that one book on my to-read list last year became part of the book club for the Committee on Women in Statistics: *Dare to Lead* by Brené Brown. I really enjoyed hearing what other statisticians and data scientists across many career stages thought of the key messages of the book.

There is a lot in our statistical training that sets us up well for leadership; we just have to know where to look to find those lessons. One of Brené Brown's key messages is that vulnerability is courageous. When we start a new project in statistics or data science, we often start with zero knowledge about the substantive issues. We need to be vulnerable and first admit that before we can make progress. Then, we ask good questions and listen carefully to learn what we must to advance the project. ■

JUAN M. LAVISTA FERRES is the vice president, chief data scientist, and lab director of the Microsoft AI for Good Lab, where he works with a team of data scientists and researchers in AI, machine learning, and statistical modeling. He joined Microsoft in 2009 to work for the Microsoft Experimentation Platform, where he designed and ran randomized control experiments across Microsoft groups. He also worked as part of the Bing Data Mining team and led a group applying data mining, machine learning, statistical modeling, and online experimentation on a large scale.

Ferres started the Microsoft efforts related to sudden infant death syndrome, and his work has been published in top academic journals, including *Pediatrics.* Additionally, his work has been covered by more than 100 news outlets around the world.

Who or what inspired you to study statistics/data science?

Two things were key. First, databases. I started to work with data when I was incredibly young and, when I learned SQL, I realized the power of answering questions with data. The second was in my algorithmic class, when I discovered the ID3 algorithm (a decision tree machine learning algorithm invented by Ross Quinlan). I became fascinated by the possibilities and power of data and machine learning.

Describe a current project and tell us what inspires you in your current work?

Working on projects that have an impact and affect the lives of others inspires me every day. We are currently working with the UN to understand the destruction of buildings in Ukraine, which are protected under the Geneva Convention. This is a collaboration with PlanetLabs, and we use deep learning on high-resolution satellite imagery.

What is one job skill you learned recently, and why do you find it important?

I have been learning and using Captum. Captum is a model interpretability and understanding library for PyTorch. Especially in deep learning, understanding the model learnings—particularly in areas like medical imaging—is a must-have. Captum is a great library that provides state-of-the-art algorithms to address this issue.

What are three pieces of advice you would offer your younger self?

Learning statistics is more important than you think. Statistics is one of the fundamental foundations of my job. As part of my computer science undergraduate degree, I took courses in statistics, but I did not pay enough attention to them because I thought they were not needed. This was a big mistake, so I had to relearn a lot of statistics years later.

Learn Python. All programming languages are good enough to work with data and develop software. From a pure programming language perspective, although Python is not the fastest or most efficient, its uniqueness is its vast community of software developers and data scientists who contribute open-source tools that provide tremendous power to data scientists.

Invest in simplicity. As humans, we need to recognize that we are addicted to complexity. We like complex projects and complicated things. This is the wrong addiction. If you want to impress people, your solutions can be complicated, but if you want to have an impact, your solutions need to be simple. Building simple solutions is hard but worth it.

What do you love most about your job?

There are especially important problems out there that can and should be solved with technology and data. I consider myself lucky because I have the opportunity to work on some of them.

In a President's Corner, Kathy offered this definition of community analytics: bringing the best of statistical science—in collaboration with municipal governments, universities, local businesses, NGOs, and community organizations—to improve lives through a better understanding of our communities and how we live, work, learn, and play. Looking to the future, what do you see as emerging areas and opportunities?

Half the world lacks access to essential health services, and there are not enough doctors to be able to provide services. Approximately 80 percent of the world's population has access to smartphones, and we expect this number to continue to increase. We predict a significant increase in possibilities to democratize health services by running telehealth, health apps, and algorithms on these devices.

What are the most important skills for students to be developing right now to be ready for the future of data science?

The fundamental skills to learn are coding and statistics. Students as young as those in middle or high school can be proficient in both. As soon as you can learn how to write and read, you can learn to code. You don't need and should not wait until you're an undergraduate to learn these skills.

Name one or two favorite blogs or books you have read and would recommend to others.

If you have to read two books, I recommend *Lectures on Probability Theory and Mathematical Statistics* by Marco Taboga and *Deep Learning with Python* by Francois Chollet. ■

TANYA MOORE is the founder of Intersecting Lines, LLC, a mission-driven company that uses analytical and community-centered approaches to empower leaders and organizations in the use of data science, statistics, and evaluation methods to support equity-focused initiatives in health, education, and workforce development. She is one of the co-founders of the Infinite Possibilities Conference, a national conference designed to support, promote, and encourage BIPOC women in mathematics and statistics. Moore has been featured in *Essence Magazine, Black Enterprise,* and *O, The Oprah Magazine* and was recognized as a "STEM Woman of the Year" by California State Assembly member Nancy Skinner.

Who or what inspired you to study statistics/data science?

What matters most to me is my family, my community, and aligning my actions with the belief that everyone deserves an opportunity to live out their dreams and share their gifts. I never thought mathematics and statistics would be the vehicle for me to create a life that allows me to actualize what I most care about, but it has.

My teachers have served as a huge source of inspiration to me. Spelman College created an environment in which to learn mathematics while in community. At Spelman, my professors believed in me, encouraged me, and challenged me. I was guided to not just think about satisfying requirements for graduation, but to build a life in mathematics that extended past college.

My high-school teacher Mr. Richard Navies had a huge influence on my understanding of the connection between history and our present-day societal challenges. He encouraged all of us to consider a life of service, using whatever skills and talents we developed to make our communities stronger and healthier.

My decision to study statistics started with my interest in analysis and probability and my desire to do work that could improve health. Biostatistics became a way to tie together the things I cared deeply about.

Describe a current project and tell us what inspires you in your current work?

The projects I'm currently working on are energizing and feel personally and professionally meaningful to me. One project that I'm supporting is focused on creating equitable health outcomes in health care systems. Multiple pilots are being launched around the country in different health care settings. What is exciting to me is that the role of evaluation is focused on what can be learned about what supports or hinders health equity, so the type of data and method of collection and analysis will be varied, contextual and organized around answering learning questions.

Another project I'm involved with is using AI to read and evaluate lots of policy documents. I've been working with a team to train the AI model in order to minimize racial and ethnic bias in the analysis.

The organizations I work with are bringing a lot of intentionality and consciousness to their work, and I'm humbled by their willingness to dive in and wrestle with some of the most challenging and complicated social issues of our time, such as racism, poverty, or generational trauma. They inspire me with their courage, brilliance, and sense of hope for the future.

I feel grateful that my training in biostatistics allows me to work across different sectors with inspiring leaders who are working toward creating a more inclusive and equitable society.

What is one job skill you learned recently, and why do you find it important?

Recently, I've been learning from one of my collaborators how to effectively use Airtable. It allows organizations to take a first step in creating a data system to organize all the data they care about before investing in high-priced software.

Many organizations that provide services or have programs in the community have data they collect—like intake forms, assessments, pre-post surveys, photos, video recordings, or written narratives—and it's typically in different file formats or housed in different places.

It's so powerful to see organizations get excited about their data coming together in one database that allows for different views and to have a system that provides greater ease in accessing actionable insights.

Most organizations these days collect a lot of data, but it often goes underutilized because it's not accessible in a way that helps teams and leaders do the sense-making and use what they learn for decision-making, program improvement, or telling their story of impact.

What are three pieces of advice you would offer your younger self?

- Stay curious. Explore and learn all that interests and excites you, even if it feels random or disconnected to your primary focus for the moment.
- Invest time in developing meaningful and authentic relationships; the world is smaller than you think.
- Be kind to yourself and forgiving of your mistakes. Learning who you are and about life happens inside and outside the classroom.

What do you love most about your job?

What I love is supporting nonprofit and foundation leaders in codifying their vision for positive social change and developing a plan of action that leads to measurable results. Through supporting their research and evaluation goals, I get a front-row seat to transformative change happening all across the country. It's brought me so much joy and a sense of hope for the future to be part of so many incredible efforts happening around the country to improve our education and health systems or that aim to strengthen and heal communities that have been underserved or marginalized.

It's also been an incredible experience to create my own company, Intersecting Lines. I get to provide services that integrate and leverage all my professional and educational skills and experiences. The entrepreneurial journey, while at times scary, provides a level of freedom, creativity, and flexibility in how and when I work.

In a President's Corner, Kathy offered this definition of community analytics: bringing the best of statistical science—in collaboration with municipal governments, universities, local businesses, NGOs, and community organizations—to improve lives through a better understanding of our communities and how we live, work, learn, and play. Looking to the future, what do you see as emerging areas and opportunities?

Improving lives through a better understanding of our communities should include community perspectives and expertise. Too often, when research or evaluation

has been conducted on communities—and even when in service to communities—it has had a way of extracting the data desired while leaving community members outside the rest of the process, excluding them from the design, analysis, meaning-making, and dissemination. I think the more we can approach community analytics with a lens of partnership and inclusivity, the more meaningful and useful the data will be.

Policy and systems change to create more equitable institutions is complex work that can take years and many partners working together to realize. I also see organizations wrestling with how to demonstrate change that is not so easily quantifiable. Addressing challenges that have their root causes anchored in racism, poverty, or trauma can feel intractable.

I think an emerging area of community analytics will be how to better integrate quantitative and qualitative data in demonstrating change over time. Just as societal issues are multifaceted, community analytics should be, too.

What are the most important skills for students to be developing right now to be ready for the future in data science?

Certainly, skills in statistics, mathematics, and computer programming are foundational to data science. Not only do they provide the tools to do the work, but studying these subjects helps to build one's muscle for problem solving and strategic thinking. Beyond that, I would encourage students to explore courses and topics outside their discipline. The more you can develop a process for learning about other topics and viewing yourself as a translator of sorts, the more effective you'll be as a data scientist.

Name one or two favorite blogs or books you have read and would recommend to others.

One exciting book I've recently discovered is *W.E.B. Du Bois's Data Portraits: Visualizing Black America,* edited by Whitney Battle-Baptise and Britt Rusert. The book showcases 60 data visualizations created by W.E.B. DuBois that were displayed in the 1900 Paris Exposition. The charts and graphs were presentations of publicly available data and collectively communicated a story of African American advances in society post-slavery. The images are a powerful example of how data can be used to not only present facts, but to shine a light on those issues in society that need our attention, awareness, and action.

I also love Stephanie Evergreen's blog (*https://stephanieevergreen.com*). She is a data viz entrepreneur whose work I've followed for a few years now. I admire her expertise; sense of humor; and commitment to making data practical, accessible, useful, and meaningful. ■

JEDI CORNER

Disabilities as Assets and Strengths

The Justice, Equality, Diversity, and Inclusion (JEDI) Outreach Group Corner is a regular component of Amstat News *in which statisticians write about and educate our community about JEDI-related matters. If you have an idea or article for the column, email the JEDI Corner manager at* jedicorner@datascijedi.org.

"My disability has opened my eyes to see my true abilities." — Robert M. Hensel

As stated by the US Centers for Disease Control and Prevention at *https://bit.ly/3pj1OGh*, "Disability impacts all of us." More than a quarter of the US population has some type of disability, many of which are invisible (e.g., chronic illness or learning disabilities) or doubly invisible (e.g., struggles with social cues). Over the past decade, societal understanding of disability has increased. Whereas the focus used to be on limitations brought on by disability, today's understanding and study of disability looks at the positive aspects disabled individuals add to their communities.

A Changing Perspective

Traditionally, in research especially, the medical model has been used to study disability. The medical model defines disability (or illness, generally) as a diagnosis or medical problem that limits the person experiencing it. The Linear Medical Model of Disability explains that, in this paradigm, an individual is defined by his/her/their disability and curing or managing the disability is believed to be achievable if a detailed clinical perspective is available. This can allow the individual to control or alter the course of his/her/their disability. While the medical model does promote the availability of resources to cure disabilities and increase functioning, it focuses heavily on individuals and defines them by their disability.

Shu-Min Liao is an assistant professor of statistics and a faculty facilitator of the "Being Human in STEM" course at Amherst College. Sparked by her disabilities, minority identities, and unusual lived experiences, Liao is passionate about STEM education research and DEI work (besides copula modeling for categorical data). She is currently part of both the JEDI Outreach Group and the ASA Committee on Statistics and Disability.

Chuck Coleman earned his PhD in economics from George Mason University and works as a mathematical statistician focusing on economic statistics. Previously, he worked as a statistician/demographer for the US Census Bureau. Coleman was diagnosed with autism in 2007. Since then, he has been an autistic self-advocate in his workplace and the community at large. He has also been working on JEDI issues in general.

Ryan Machtmes, GStat, is a person with partial blindness and a lifelong advocate for people with disabilities. He is a longtime member and former chair of the Committee on Statistics and Disability and is currently a fellow in the American Foundation for the Blind's Blind Leaders Development Program. Machtmes holds a master's degree in applied statistics from Louisiana State University and is a former presidential management fellow.

When **Peter Flom** was 5 years old, a psychologist told his parents he had "minimal brain dysfunction" and "would never go to college." Flom skipped two grades and graduated from college at age 20. Eventually, he earned a PhD in psychometrics and worked for 25 years as a statistician. He has written a book, *Screwed Up Somehow but Not Stupid: Life with a Learning Disability,* and has a website at *www.IAmLearningDisabled.com*.

Erin Chapman works on Amazon AWS's cryptographic computing team. With degrees in mathematics and computer science, she's spent her career working in various intersections of the two fields. Her service dog, Valor, is a constant at her side. She spends her free time advocating for DEI. She also volunteers with a variety of educational organizations. When not working or volunteering, you can often find her enjoying the outdoors with her family.

Anja Zgodic is a doctoral student in the department of biostatistics at the University of South Carolina who previously worked as a data scientist in industry. Her research interests are in longitudinal data analysis, high-dimensional data analysis, and Bayesian statistics. Zgodic is the current chair of the ASA Committee on Statistics and Disability.

Alternatively, The Social Model of Disability removes the focus from the individual experiencing disability and centers it on systemic barriers and exclusions in society as factors that disable individuals. According to this newer framework established by the World Health Organization, it is really the failure to design society as accessible to individuals with any type of impairment that leads to disability, even while various physical, sensory, intellectual, or psychological factors may cause impairments to individuals.

In the medical paradigm, People with Disabilities (PWD) are considered patients who can receive an intervention to be cured of their disability, or even research subjects to find a cure (which has left a certain unfortunate legacy, as evidenced by human subjects training). However, in the social model of disability, PWD are active community members with agency to participate fully in decision-making processes that remove barriers and increase accessibility.

People with Disabilities Are Assets

Despite human imperfection, many people innately have a disparity mindset that highlights differences. However, more and more studies reveal benefits and advantages of asset-based thinking (focusing on strengths and what is right) over deficit-based thinking (focusing on weakness or what is wrong). In the context of disability, this research suggests society should shift from the medical model (deficit-based thinking) to the social model (asset-based thinking).

The challenges and lived experience of disability foster the development of many unique skills and talents, which contribute significantly to workplace success among PWD. These unique skills and talents include creativity, persistence, resilience, and problem-solving, which are developed as a result of the necessity of finding alternative ways to complete tasks under limitations. Moreover, the struggles and frustrations of living with disability also supply PWD with distinct perspectives and increased compassion and empathy for others around them, which enable disabled people to strengthen teamwork and communication in the workplace.

Many PWD are enthusiastic advocates for diversity, equity, and inclusivity (DEI)—a must for a healthy, productive, and thriving workplace. As Ilse Daehn and Paula Croxson point out in a September 2021 *Science* letter, people with disabilities contribute in powerful ways to the success and productivity of workplaces:

> Now is the time to acknowledge that individuals with disabilities are strong and innovative contributors to society. By valuing traits such as compassion, empathy, and humility, we can empower all people in STEM, making our field better, not weaker.

People with Disabilities Strengthen Everything

1. People with Disabilities Improve Workplaces

Disabled employees improve workplaces when they are well-accepted and provided reasonable accommodations (as mandated of employers by federal civil rights law). PWD contribute new ideas to help drive innovation by sharing unique experiences and perspectives.

According to a September 3, 2019, *New York Post* article titled "Workers with Disabilities Bring a Range of Strengths and Assets to the Job," "a recent study from Accenture, in partnership with Disability: IN and the American Association of People with Disabilities, found that companies that make efforts to hire those with disabilities performed better and saw, on average, 23 percent higher revenue."

This may explain why most large companies, like Microsoft and Google, have dedicated program managers for improvement of accessibility and always try to find ways to recruit more and better support PWD in their companies.

2. Accommodations for Disabilities Drive Innovation

Have you heard of the "curb-cut effect"? Angela Glover Blackwell, in a 2019 *Stanford Social Innovation Review* article, describes it as a phenomenon of how "laws and programs designed to benefit vulnerable groups, such as the disabled or people of color, often end up benefiting all of society." It should not be a surprise that this term came from the design of curb cuts, which were originally invented for wheelchair users, but parents with strollers or travelers with wheeled suitcases are also benefiting from this design.

The nonautistic "neurotypical" style of cognition is top-down: One sees the grass before the blades. In contrast, the autistic style of cognition is bottom-up: One sees the blades before the grass.

As another example, electric toothbrushes were initially created to help people with limited mobility and control do a better job of brushing, but many studies soon found that all people's teeth and gums can benefit from such a design.

It might be obvious that audiobooks were first developed for vision-impaired readers, but now everybody can enjoy the convenience of those products.

Find more examples in "8 Everyday Items Originally Invented for People with Disabilities" at *https://science.howstuffworks.com/innovation/everyday-innovations/items-invented-people-with-disabilities.htm*.

3. Disabilities Advance Education

The aforementioned innovations are good examples of universal design (UD), which, according to the University of Washington's DO-IT (Disabilities, Opportunities, Internetworking, and Technology) group, is a process of "creating products, buildings, or environments to accommodate for a wide range of abilities and disabilities, making them accessible and beneficial to *all* users."

UD is an aggressive and proactive approach to accommodating all individuals (but does not obviate all accommodations). If we extend this concept to education and intentionally make course materials, practice, and environments accessible and engageable to *all* students, such a framework is called universal design for learning (UDL). UDL is known to promote JEDI, support culturally responsive teaching, and further transform student learning.

4. Disabilities Inform Research

Autism is an example of the changing paradigm of disability, from defect to valuable difference. From The Autism History Project, we learn that Eugen Bleuler provided the initial definitions of autism and schizophrenia in 1911, with autism being the most severe type of schizophrenia. Being a proponent of eugenics, he argued that "schizophrenics" should not be allowed to reproduce.

However, as the work of Hans Asperger became more widely known in the 1990s, autism came to be understood as a different style of cognition. The nonautistic "neurotypical" style of cognition is top-down: One sees the grass before the blades. In contrast, the autistic style of cognition is bottom-up: One sees the blades before the grass.

Pattern recognition is very strong. Autistics are more interested in the ground truth, rather than its representations. Autistic traits exhibit higher variances. Both hyperlexia—extreme facility with languages—and language difficulties—including intermittent or permanent inability to speak—are common, sometimes in the same person.

The combination of attention to detail, pattern recognition, interest in ground truth, and other enhanced abilities are of great value in research. Autistics are overrepresented in STEM and perhaps other occupations. Autism is an example of how the loss of ability, mainly social in this case, enhances other abilities. By removing barriers, all of society benefits from the contributions of PWD.

We, as the Committee on Statistics and Disability—a member committee of the American Statistical Association's Membership Committee and a partner with the JEDI Outreach Group—advocate for ASA members with disabilities. Recently, we created a new committee charge, which adopts for our work the social model for disability and transitions away from the medical model. While we support the advancement of medical science and do not deny its successes, we recognize that disability is also an innate and immutable factor of social difference equal to those of race and gender expression and must also be reconciled socially. Thus, we advocate for the removal of systemic inequities that hinder people with disabilities from full access and participation within the statistics profession. ■

MORE ONLINE
View additional links related to this article on *Amstat News* online: *https://magazine.amstat.org.*

STATS4GOOD

ASA STUDENT PROGRAMS

Create D4G Experiences, Opportunities

David Corliss is lead, Industrial Business Analytics, and manager, Data Science Center of Excellence, Stellantis. He serves on the steering committee for the Conference on Statistical Practice and is the founder of Peace-Work.

With the start of a new school year, Stats4Good takes a look at some of the ways the ASA empowers students to act for the greater good.

One of the most important ways the ASA supports D4G work by students is through its more than 100 student chapters, including three in foreign countries. The chapters help students network, bring in speakers, write and present research, and learn about internships and job opportunities. They also make the perfect setting for Data for Good projects.

For example, the student chapters at George Mason and George Washington universities organized a hackathon in 2018 to identify informative features in the Global Terrorism Database.

No chapter at your school? No problem—the ASA can help. Ask a faculty adviser or chapter president to fill out the application form at *https://bit.ly/3JOmaRo.*

Regular ASA chapters can get involved, too. Having a speaker talk about their D4G project is a great way to reach out to new members and even the general public. Chapters also support education in math and statistics, which is one of the Data for Good community's most important activities, and promote D4G by judging science

The chapters help students network, bring in speakers, write and present research, and learn about internships and job opportunities.

Get Involved

In opportunities this month, the International Conference on Health Policy Statistics is accepting applications for student travel awards through September 15.

Statistical advocacy through policy is one of the most effective ways to turn science research into major public impact. Visit *ww2.amstat.org/meetings/ichps/2023/travelawards.cfm* for details.

Also, the start of the academic year is a great time to get involved. Mentor a student, be a statistical expert for a capstone project, volunteer for DataFest, help with high-school science fair projects … there are so many ways to be an advocate!

Like all of statistical science, Data for Good is only a generation or two away from extinction. By involving and supporting students, we can all help build a bright future for Data for Good.

MORE ONLINE
Learn about student chapters at *www.amstat.org/membership/asa-communities/student-chapters.*

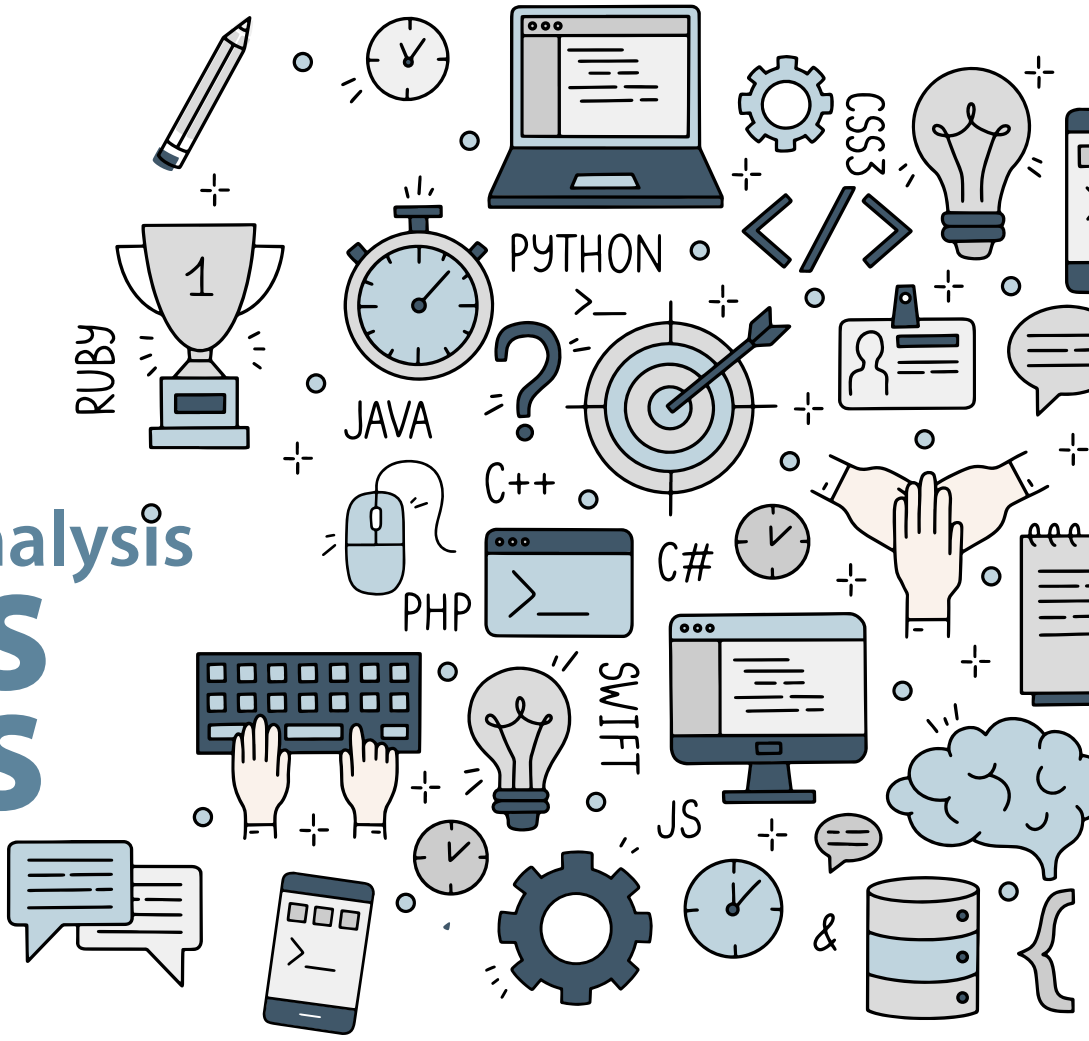
fairs and participating in hackathons. As the leading advocate for statistical science in their communities, ASA chapters should know best the particular needs in their local area and facilitate the involvement of experts to promote data-driven solutions.

Another way the ASA supports student activities in D4G is through committees and outreach groups. While this will be especially important for D4G-oriented teams like JEDI or science policy, nearly all ASA committees and outreach groups have a strong student component.

Students also play a vital role in ASA conferences. Student posters are one of the best ways to get more actively involved in the ASA, learn presentation skills, and network with a variety of people in your field. Student posters this year at JSM included research from dozens of students in every area of Data for Good.

Nearly every ASA program creates opportunities for student, and they can all be harnessed for Data for Good projects. Whether you are a student, educator, or mentor, there are opportunities to be involved in making an impact for good today while building the future of Data for Good.

You can always email me at *davidjcorliss@gmail.com* to get information about D4G activities or let me know about what you have been doing. ■



STAT*tr@k*

Statistical Analysis SOLVES CRIMES

In 2014, Italian nurse **Daniela Poggiali** was arrested and convicted of murdering two hospital patients. Richard Gill, a statistician in the United Kingdom, followed the case and became suspicious of the analysis used to convict her. Gill enlisted the help of Italian forensic statistician and ASA member **Julia Mortera** and, together, they secured an acquittal for Poggiali after finding errors in the original statistical analysis.

We wanted to know more about what it takes to become a forensic statistician, so we asked Mortera to answer the following questions:

Julia Mortera is a statistics professor at the University Roma Tre and honorary professor at the University of Bristol. She is also a member of the Royal Statistical Society Statistics and Law Section and the ASA Forensic Science Committee. The editor of *Law, Probability, and Risk*, she has made contributions to forensic statistics, decision theory, and Bayesian statistics.

What motivated you to become a forensic statistician?

It was quite by chance. I was asked to discuss a paper on the "island problem" that Phil Dawid had written about. The island problem is a toy example that illustrates the uses and misuses of statistical logic in forensic identification. I then worked with Phil on some generalizations of the problem and the results were published in a 1996 *Journal of the Royal Statistical Society B* joint paper, "Coherent Analysis of Forensic Identification Evidence."

I so enjoyed the statistical problems that arose in the area of forensic statistics that I have never left the field in more than 25 years! I had some genetics background, having previously worked with geneticists on medical problems, which led my focus to forensic genetics problems.

Describe your path to becoming a forensic statistician.

My career in the area was riddled with difficulty as, being a new field in Italy, it was not recognized by statisticians, which implied that papers in the area were not considered valid for promotion. This is now changing, but it took a lot of effort to obtain this recognition. This means only a few statisticians in Italy have become involved in this exciting area, which is a great pity.

What do you like most about your work in this area?

I really enjoy the interdisciplinary aspect of the area, as it brings together diverse fields such as statistics, genetics, and law, to name a few. It is an exciting field, as new and interesting methodological issues can arise from casework data and technology moves forward. I also enjoy developing probabilistic expert systems models, which have proved to be fundamental for helping to solve complex problems in forensic genetics.

What are the main challenges of being a forensic statistician?

There are many challenges in this field. One of the main ones is trying to communicate what I believe is the logical way to analyze and interpret evidence to forensic analysts and in the legal setting in an easily understandable way.

Understanding that you and Gill described your research on the Daniela Poggiali case in a journal article currently under review (*https://arxiv.org/pdf/2202.08895.pdf*), would you provide a short summary that includes the statistical aspects for our readers?

Suspicions about medical murder sometimes arise due to a surprising or unexpected series of events, such as an apparently unusual number of deaths among patients under the care of a particular nurse. There is a statistical challenge of distinguishing event clusters that arise from criminal acts from those that arise coincidentally from other causes.

In the Poggiali case, we examined some of the possible confounders that could explain away the higher death rates when Daniela was on duty.

Simple exploratory data analysis revealed interesting facts. For example, a death registered to have occurred on a particular day is associated with a nurse, even if she is on duty for just a fraction of that day. The actual time of death differs from the recorded time of death. Deaths are recorded more often at 7 a.m. (in the overlap between night and morning shifts), at midnight, and on the hour or half hour. So, a nurse like Daniela who starts her morning shifts early and finishes her night shifts late will be associated with more deaths during hours she was working than hours she was not working.

The prosecution stated the rate of deaths dropped after Daniela was dismissed from the hospital. However, a simple analysis showed admissions dropped considerably then (the hospital had become infamous due to media coverage) and, consequently, deaths diminished!

So, a difference in the mortality rates between different nurses could be due to confounding variables. Even if all the measured confounders are taken into account, no causal effect can be concluded from a mere association in an observational analysis like this.

Later in the case, the defense showed me the analysis based on a simple linear regression between postmortem interval and "vitreous humor" potassium concentration made by the prosecution's forensic pathologist, who declared potassium chloride poisoning was the cause of a patient's death. I realized that in making his prediction of postmortem vitreous humor potassium concentration, he did not consider a prediction interval and did not take into account any uncertainty or variability due to various factors. Considering this, even if the concentration observed at the time of postmortem was above average, it was well within a 95 percent prediction interval.

Causal effects cannot be inferred by just providing a descriptive analysis, which can reveal—at most—an association among the presence of a given nurse and an increased mortality rate. This association is due to many confounding factors that, if not accounted for appropriately, can lead to a misleading conclusion.

We showed how the results of a generalized linear model were in contrast to the findings in the report by the prosecution's expert witnesses. Using the model, in fact, we showed the increase in the number of deaths had no relation to Daniela's presence.

A take-home message of the case is association is not causation!

MORE ONLINE

Read more about the Poggiali case at *https://bit.ly/3v2Dvjw.*

For details about the statistical analyses used in the Poggiali case, visit *https://arxiv.org/pdf/2202.08895.pdf.*

Had you ever worked on a case like Poggiali's before? If so, tell us a little about it.

I have not worked on similar cases. There are luckily not too many in Italy. I have worked in the investigative phase of several problems concerning DNA mixtures. Recently, I worked on a missing persons case in which the evidence was a DNA mixture presumably from related contributors. Another case concerned the search for the culprit of a brutal murder of an 11-year-old in which, again, the only evidence were DNA mixtures found on her clothing. I also worked on an incomplete paternity recognition of a deceased famous Italian singer in which the evidence consisted of a mixture of DNA.

Essential to success in this area is identifying a statistical flaw in the legal reasoning and then communicating the issue in a way that is understandable to lay people. How do you address this challenge?

I have been involved with a group of statisticians and law scholars for the RSS Statistics and the Law Section on *Statistical Issues in Investigation of Suspected Medical Misconduct,* which provides advice and guidance on the investigation and evaluation of such cases. This report was prompted by concerns about the statistical challenges such cases pose for the legal system, since statistical evidence is difficult for lay people and even legal professionals to evaluate.

I think it is important to train young students—researchers in legal studies—how to interpret and understand probabilistic-statistical reasoning in court cases. I have been teaching groups of legal students and practitioners, forensic scientists, and investigative police the basic principles underlying our field. Communication and finding a common understandable language to portray statistical concepts, even in complex settings, is a challenge but vital to undertake.

What advice do you have for students and early-career statisticians who might pursue forensic statistics?

I'd like to tell them the following:

- This is an exciting area of research
- The applications can bring on new interesting methodological issues
- The interplay of different disciplines makes it a challenging and thought-provoking area
- They should never take anything for granted in this field, but always go back to first principles and look out for the pitfalls in legal and forensic reasoning
- Most importantly, they can help deliver justice ■

Ethel Newbold Prize

The Bernoulli Society's Newbold Prize Committee invites nominations for the fourth Ethel Newbold Prize, which is awarded biannually to an outstanding statistical scientist in early or mid-career for work that represents excellence in research in mathematical statistics and/or excellence in research that links developments in a substantive field to new advances in statistics.

Ethel May Newbold (1882–1933) was an English statistician and the first woman to be awarded the Guy Medal in Silver by the Royal Statistical Society, in 1928. During her academic career (1921–1930), she published 17 papers in statistics and subject-matter journals. Read more about Newbold in the *Journal of the Royal Statistical Society* from 1933 at *www.jstor.org/stable/2341811*.

The name of the prize recognizes a historically important role of women in statistics. The prize, itself, is for excellence in statistics without reference to the gender of the recipient. In any year in which the award is due, the prize will not be awarded unless the set of all nominations includes candidates from both genders.

The award consists of a cash prize of 2,500 euros and a certificate. The awardee will be invited to present a talk at a Bernoulli World Congress, Bernoulli-sponsored major conference, or ISI World Statistics Congress.

Each nomination should include a letter outlining the case in support of the nominee, along with a curriculum vitae. Nominations, as well as any inquiries about the award, should be sent to Gesine Reinert at *reinert@stats.ox.ac.uk.* The deadline for nominations is November 30; the prize winner will be selected in the spring of 2023.

Newbold Prize Committee Members: Gesine Reinert, chair; Adrian Röllin and Susan Murphy

Symposium Focuses on Opportunities for Massachusetts Community Colleges

Benjamin S. Baumer, Nicholas J. Horton, Ethan Meyers, and Andrea Dustin

The NSF-funded Data Science Corps Wrangle-Analyze-Visualize (DSC-WAV) project organized a symposium of academic leaders, faculty, and other stakeholders to foster the development of flexible and inclusive data science pathways in Massachusetts. The symposium was hosted by Smith College on June 13 with the following two goals:

- Advocate for the creation of data science transfer pathways
- Identify barriers and opportunities for student success in data science

Participants focused on the following questions:

- Where is the passion to make data science skills and capacities available to your students?
- What questions will these students be able to answer about their local community and their lives?
- How can we help students build these transferable skills while also making an impact locally?

Marc Maier—head of data science, risk, and product at MassMutual—opened the symposium with an address in which he described how data science has transformed the company. He shared how the 200 data scientists, engineers, and technologists at MassMutual have worked to improve decision-making at all levels of the company. He also noted how the company's focus on drawing from a diverse pool of applicants and creating an inclusive workplace culture within the data science development program has resulted in a team that is 50 percent non-white and 70 percent non-male.

Chief scientist and associate director for research and methodology at the US Census Bureau Sallie Keller's keynote address highlighted the importance of community colleges in her academic trajectory. Through the lens of her current work, she offered numerous examples of how data can be used to inform policy decisions in many realms. Examples included resource allocations in cities and ways to address disparities in services.

A panel with speakers from Bunker Hill Community College, Roxbury Community College, and the University of Massachusetts, Amherst shared innovative programs in Massachusetts, including a new data analytics program, a new certificate, and efforts to support transfer students at four-year institutions.

The 30 symposium participants came from 13 institutions and organizations. They spent time together in small groups identifying strengths, weaknesses, opportunities, and threats for their institutions and students and worked to identify next steps to overcome barriers. Ideas from these discussions were incorporated into a white paper that identifies five curricular friction points that complicate transfer pathways.

The symposium leaders' plans are to develop a roadmap to make data science opportunities available to all two-year college students in Massachusetts and beyond. ■

MORE ONLINE
More information about the symposium and resources related to the project, including the white paper, can be found at *https://dsc-wav.github.io/www/outreach.html*.

Virtual Conference to Celebrate Women in Statistics, Data Science

Tomi Mori, Vanda Lourenço, Miguel de Carvalho, Altea Lorenzo-Arribas, Jessica Kohlschmidt, and Donna LaLonde

The Caucus for Women in Statistics and Portuguese Statistical Association will host the **International Day for Women in Statistics and Data Science** October 11 to celebrate female statisticians and data scientists around the world.

The aims of the virtual conference are to:

- Showcase women and their contributions to the field
- Connect women statisticians and data scientists around the world
- Encourage collaborations among statistical societies around the world
- Prompt statistics and data science to become more inclusive and diverse
- Bridge statistics and data science

There will be both live and recorded presentations. Submit session ideas to *idwsds1@gmail.com*. For more information, visit *idwsds.org* or follow updates on Twitter by searching for @cwstat.

Finding the De-Anonymization Needle in the SEER Haystack

Chris Barker, Statistical Consulting Section Chair-Elect

As chair-elect of the Statistical Consulting Section, I present my motivation for one of my several forthcoming section initiatives. My initiative arises because I recently needed a crash course in concepts entirely new to me about data privacy, anonymization/de-anonymization, identification/de-identification and re-identification, and statistical disclosure. Based on what I learned in my crash course, I am inviting interested statisticians to help develop a "data privacy toolbox" that members of the consulting section and certainly any/all statisticians at the ASA can use in their day-to-day work. The toolbox may be used at a leisurely pace, rather than as a crash course. Volunteers need not be section members, though I encourage joining the section.

Defining and measuring the success of the toolbox is an additional objective for the group.

Privacy in the 21st century may no longer exist. Bill Gates stated in a 2013 *Wired Magazine* article, "Historically, privacy was almost implicit, because it was hard to find and gather information." Today, de-anonymization (i.e., re-identification)—the practice and relative ease of 'adversaries' matching anonymized data (i.e., de-identified data) with publicly available, or auxiliary data—may lead to identifying that "anonymous" person in terms of actual name, address, employer, etc.

The ethics of de-anonymization and its implications for those de-anonymized has been studied by independent ethicists and other experts, including those at the US Census Bureau. One data use ethics issue is the unambiguous violation of explicitly stated terms of use for the Surveillance and Epidemiology End Results (SEER) and clinicaltrials.gov (CTG) databases.

The specific paper, "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials" in the *American Economic Review* by Eric Budish, Benjamin Roin, and Heidi Williams, clearly states the researchers and organizations linked SEER and CTG with no reference to the terms of use (*https://seer.cancer.gov/data-software/documentation/seerstat/nov2021/seer-dua-nov2021.html*). The authors, as well as Nobel Prize–winning journal editor Esther Duflo and Nobel Prize–winning president of the American Economic Association David Card, when asked, did not provide proof that the authors had permission of any kind from the federal agencies overseeing SEER and CTG to violate the terms of use. This creates a risk that adversaries will re-identify oncology patients by linking to auxiliary data.

Statisticians working with any data from humans may need to update their understanding of anonymization, de-anonymization, and statistical disclosure.

The Needle in the SEER Haystack

Paraphrasing Bruce Schneier in his *Wired* article, "Why 'Anonymous' Data Sometimes Isn't," what we have learned about anonymization of patient data and "statistical disclosure" may be completely outdated or possibly wrong.

SEER data is "anonymized" and information permitting identification (name, address, credit cards, salary, etc.) of individual patients has been removed. However, in 2008, Netflix created the Netflix Prize and provided anonymized customer data to the public. According to Arvind Narayanan and Vitaly Shmatikov in their *IEEE Symposium on Security and Privacy* paper, "Robust De-Anonymization of Large Sparse Datasets," a team of computer scientists was able to link the Netflix database with Internet Movie Database (IMDb), identify the Netflix

client's actual name and address, and receive confirmation of correctly identifying the individuals from Netflix management.

A critical caveat emptor to my work here is there is no direct proof of a de-anonymization, since that can only be achieved by directly contacting the patients involved. Briefly, I inspected (using SAS and R) data sets prepared by the authors and available for download by anyone with internet access. No password is required and there is no method to track the download. I found 40 unique clinical trials with exactly one patient (sample size n=1) linked to SEER patient-level data with a large number of covariates that can be used by an adversary to link to auxiliary data for de-anonymization. I have no way to contact the individual patients, and I turned over my discovery to the experts at SEER and CTG. The patient data I found is at potentially very high risk of de-anonymization. Given the detailed data available, it may be possible to de-anonymize the 40 unique cases of patients in clinical trials n=1.

I specifically asked the authors, Duflo, and Card to carry out the turnover to be compliant with the SEER terms of use to notify SEER of de-anonymizations. In the absence of their replies, I intervened—guided by the ethical principles of the ASA—and reported the matter to SEER and CTG.

As a courtesy, I specifically informed the Division of Cancer Control and Population Sciences, National Library of Medicine director, and privacy experts at SEER and CTG that I did not expect or need to know how the matter was handled.

Proof of Concept of De-Anonymization of SEER Using Certain CTG Trials

My background is in pharmaceutical clinical trials, where we routinely blind patients, investigators, and sponsors. Anonymization and de-anonymization differ from blinding and unblinding. The two share a common characteristic in that individual patient identifiers are removed by an anonymization algorithm, sometimes referred to in the privacy literature as "catch and release." The two differ in that de-anonymization may occur only for a single patient, several patients, or possibly all patients. Clinical trial unblinding is applied to all patient data at one time, called a "database unlock." The patient identifiers are not included in journal publications. Protection of pharmaceutical clinical trials patient data is addressed by the European Medicines Agency (2019) and European General Data Protection Regulation.

I believe I have discovered the first-ever publication of the "proof of concept" for a de-anonymization algorithm in a prominent peer-reviewed journal of economics policy, the *American Economic Review.* The proof of concept of de-anonymization of SEER anonymized data is caused by combining two crown jewel data sets of the federal statistical system—SEER and CTG—in violation of the terms of use. In fact, as many as four federal-level databases may be involved.

Clinicaltrials.gov, in a small number of situations, has a type of "patient-level" data, specifically a small number of clinical trials in which the final sample size (number of patients) is one (n=1). Based on my experiences, I pose broader questions that I do not attempt to answer. How important is the discovery of a proof of concept to the ongoing initiatives by pharmaceutical companies to provide data sets of pristine anonymized clinical trial data to external experts? And does the proof of concept increase the risk that some of those experts may attempt to de-anonymize the data? Last, does the proof of concept scale up to larger clinical trials?

I completely turned over the matter to the Division of Cancer Control and Population Sciences/SEER and National Library of Medicine/CTG for the privacy experts to address.

Synergism with Other ASA Committees

At the outset, I recognized the concept of a data privacy toolbox might overlap with the work of other ASA committees. To avoid duplication of effort, I invite Statistical Consulting Section members and members of ASA committees such as Data Privacy, Record Linkage, Epidemiology, and Ethics to collaborate on this initiative.

For information, contact me, Chris Barker, at *chrismbarker@yahoo.com*. ■

Professional Opportunities vacancies will be published on the ASA's website (*www.amstat.org*). Vacancy listings will appear on the website for the entire calendar month. Ads may not be placed for publication in the magazine only; all ads will be published both electronically and in print.

These listings and additional information about the 65-word ads can be found at *ww2.amstat.org/ads.*

Employers are expected to acknowledge all responses resulting from publication of their ads. Personnel advertising is accepted with the understanding that the advertiser does not discriminate among applicants on the basis of race, sex, religion, age, color, national origin, handicap, or sexual orientation.

Also, look for job ads on the ASA website at *https://jobs.amstat.org/jobseekers.*

Indiana

■ Faculty positions (rank commensurate with experience/qualifications), Department of Biostatistics/Indiana University School of Medicine, Indianapolis, IN. Duties: statistical research, teaching, collaborative research. PhD in biostatistics, statistics or related field, excellent communication skills required; Practical experience preferred. Competitive salary/excellent benefits. Submit CV, research/ teaching statements, 3 references to: *https://indiana.peopleadmin.com/postings/12607*. Indiana University is an EEO/AA employer, M/F/D/V.

Iowa

■ The Department of Statistics and Actuarial Science at the University of Iowa invites applications for a tenure-track faculty position in statistics at the rank of Assistant Professor starting August 2023. Candidates should submit applications online by November 15, 2022, at *https://jobs.uiowa.edu/faculty/view/74523*. PhD in statistics or related area is required. Iowa is an equal opportunity/affirmative action employer.

■ The Department of Statistics and Actuarial Science at the University of Iowa invites applications for an instructional-track faculty position in statistics at the rank of Lecturer starting August 2023. Candidates should submit applications online by November 15, 2022, at *https://jobs.uiowa.edu/faculty/view/74528*. Master's degree in statistics or related area is required; PhD is preferred. Iowa is an equal opportunity/affirmative action employer.

New York

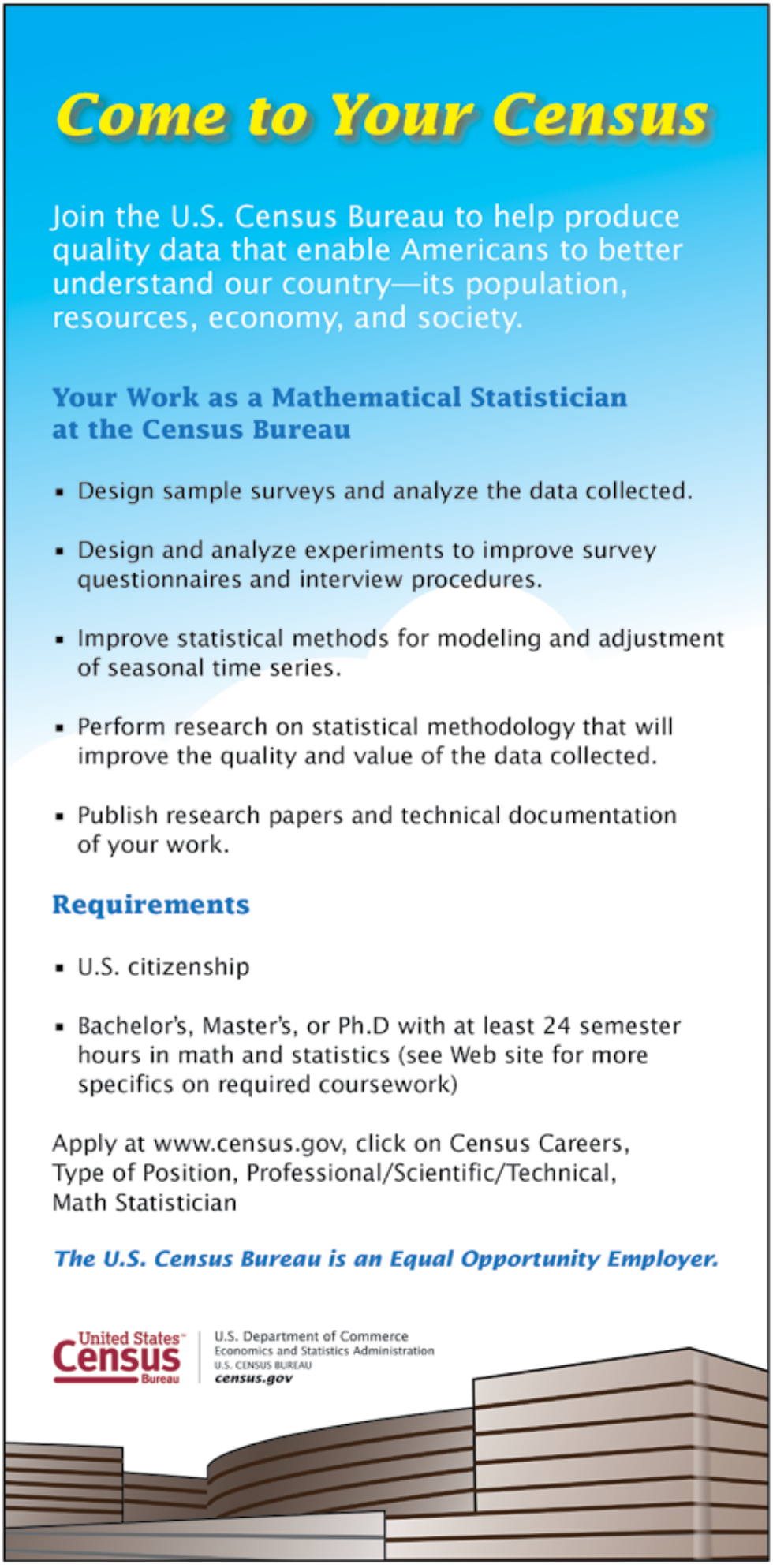
■ The Mount Sinai Institute for Health Equity Research (IHER) is seeking a skilled and motivated Director for IHER's Data Science Core who will contribute actively to addressing disparities in health and healthcare through research, leadership, and mentorship. PhD in Data Science, Epidemiology, Biostatistics or Computer Science and a at least 10 years of research experience is required. For more information visit: *https://jobs.amstat.org/jobs/17234657/senior-faculty-position-director-data-science-core.* ■

ASSISTANT PROFESSOR IN THE APPOINTMENT STREAM

The Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh invites applications for a full-time faculty position at the level of Assistant Professor in the appointment stream. The position is available immediately and requires a doctoral degree in biostatistics, statistics, epidemiology, data science, bioinformatics, or a related field with experience in managing and analyzing data from clinical research studies, including randomized clinical trials and observational studies. The successful candidate will be part of a research group involved in designing, coordinating, and analyzing clinical trials and epidemiologic studies. The individual would be expected to participate in study design, study management, data analysis, supervising students or staff, preparing data reports, and writing manuscripts. This individual will also be expected to assist with teaching, by lecturing in courses, and mentoring students. This position is funded by grants from the National Institutes of Health and other funding organizations. Salary will be commensurate with experience.

Review of applications will commence upon receipt of all application materials and will continue until the position is filled. Please apply by going to www.join.pitt.edu and applying for requisition #22001779. Please attach a cover letter, curriculum vitae, a statement of current and future research directions, and the names of three references to your online application.

The University of Pittsburgh is an Affirmative Action/ Equal Opportunity Employer and values equality of opportunity, human dignity and diversity, EOE, including disability/vets.



AMSTATNEWS

ADVERTISING DIRECTORY

Listed below are our display advertisements only. If you are looking for job-placement ads, please see the professional opportunities section. For more job listings or more information about advertising, please visit *www.amstat.org*.

misc. products and services

professional opportunities

software

This month's Top 10 is the '**Top Ten Movie Titles Ruined by Adding a Statistical Word or Phrase**.'

Wasserstein

Amstat News continues its hilarious offering by ASA Executive Director Ron Wasserstein. Each month, he delivers a special Top 10—one that aired during a recent edition of the *Practical Significance* podcast. This month's Top 10 is the "Top Ten Movie Titles Ruined by Adding a Statistical Word or Phrase."

10

Regression to the Mean Girls

09

It's a Wonderful Life Table

08

Young Frankenstein Shrinkage

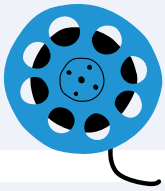
To listen to the *Practical Significance* podcast, visit *https://magazine.amstat.org/podcast-2*.

07

The Texas Markov Chainsaw Massacre

06

Das Bootstrap



05

The Fast Fourier Transform and the Furious

04

Gross Pointe Estimate Blank

03

Random Walk the Line (I could also have gone with Random Forrest Gump here.)

02

Rebel Without a Causal Inference

#01

2001: A Sample Space Odyssey

WHAT'S EASY TO SEE IN OUR CLOUD? OPPORTUNITY.

It's easier than ever to access advanced cloud analytics with SAS® Viya® on Microsoft Azure. Viya is optimized for Azure, which means it's easier to find, share and analyze information all in one place. So more people can ask the questions that lead to more opportunities. Together, SAS and Microsoft make it easier to find answers in the cloud.

Learn how at sas.com/microsoft

sas | Microsoft