
**Data Science Community Newsletter** features journalism, research papers and tools/software for July 16, 2021.

**Please let us** (Micaela Parker, Steve Van Tuyl, Catherine Cramer, Brad Stenger) know if you have something to add to next week's newsletter. We are grateful for the generous financial support from the Academic Data Science Alliance.

## UPDATES ON DATA GUARDIANSHIP

One of the approaches to ethical data science that has been promising and has, therefore, received productive attention is the idea that scientific, demographic, attitudinal, and behavioral data could and should be held in public trusts. The argument is that when more groups can access data, they'll come up with a slew of socially beneficial use cases that the company or entity which originally gathered the data would not, because any given entity has a limited purview. With thoughtful bumper guards to protect and empower data subjects, data trusts could become a bigger part of core sociotechnical infrastructure, though they are not new. Arguably, the **U.S. Census Bureau**, **National Oceanic and Atmospheric Administration**, and the **Department of Education** already act like data trusts. If you're new to this concept or want to include data trusts on your syllabus, there's an entry-level **explainer/op-ed here**.

In a more advanced discussion of data privacy, reporter **Issie Lapowsky** dug through **World Wide Web Consortium** (W3C) forum white papers and comments **on GitHub** in an attempt to decipher how big tech companies, academics, open web advocates, and regulators use this "niche" open web infrastructure group to set the terms of engagement for web browsers, network security, and a host of other technical network standards. Because participating in the W3C costs some money and a lot of time, the biggest tech companies – **Google**, **Apple**, **Microsoft** – tend to guide discussion. The work and opinion coming through the W3C is highly influential and, according to **Lapowsky's analysis**, was a contributing factor to the UK's **Competition and Markets Authority** deciding to take an official review of **Google's Privacy Sandbox project**, which recently introduced delays (possibly to accommodate the CMA's review). The takeaway here is that the technical specs surrounding "privacy" – which has a lot to do with tracking people's online and mobile phone behavior – are still in development in a forum that keeps notes transparently, isn't always friendly (hi, hi, hi – open source isn't always friendly), has a lot of big tech players, but also admits people and perspectives from smaller companies, academia, and think tanks. The Privacy Sandbox project would eliminate third party cookies from being stored on the Chrome browser – "the third party cookies are the main mechanism by which users are tracked across the web. We eventually need to remove that functionality, but we need to do it in a responsible manner" – royally pissing off small and mid-sized adtech companies, one of which raised the initial complaint with the UK CMA that seems to have delayed the Privacy Sandbox project for at least a year. Consumers, studies show, are generally in favor of setting their tracking preferences in a browser one time, a technical approach that has been attempted, but could not come to effective fruition under the then-current

regulatory and big-tech power broker environment.

If you like talking about where business models meet tech ethics issues, you are welcome to join an impromptu book club to use the new book **"An Ugly Truth: Inside Facebook's Battle for Domination"** by reporters **Sheera Frenkel** and **Cecilia Kang** as a jumping off point. It's based on 400 interviews with people who work or worked at **Facebook**. One interviewee, **Yael Eisenstat**, recorded **a podcast interview** back in 2019. Email me (**laura.noren@nyu.edu**) or go to the ADSA Slack general channel to join the book reading group – it's just two of us so far.



## HUMAN-MEDICAL INTERACTION

Medical students are getting introduced to artificial intelligence tools, and to the pros and cons of using them in their practice, as **highlighted** at a recent meeting of the **American Medical Association**. The **World Health Organization** (WHO) released **the first global report on AI in healthcare**, based on the work conducted over two years by a panel appointed by WHO. “Like all new technology, artificial intelligence holds enormous potential for improving the health of millions of people around the world, but like all technology it can also be misused and cause harm,” said **Tedros Adhanom Ghebreyesus**, WHO director-general. The report clearly states the danger of a growing reliance on AI development over core investments and strategies that are necessary to achieve universal health coverage, subordinating the rights and interests of patients and communities to commercial interests of tech companies or governmental policy, e.g. “unethical collection and use of health data; biases encoded in algorithms, and risks of AI to patient safety, cybersecurity, and the environment.” The promise of AI is not free; it comes with ethics and human rights obligations at every stage of the design, development, and deployment process. WHO’s report includes six principles for AI regulation and governance: 1) Protecting human autonomy; 2) Promoting human well-being and safety and the public interest; 3) Ensuring transparency, explainability and intelligibility; 4) Fostering responsibility and accountability; 5) Ensuring inclusiveness and equity; 6) Promoting AI that is responsive and sustainable.

A new computer model offers an **automated tool** to study serious medical conversations in large, inclusive, and multi-site epidemiological studies. Developed at the **University of Vermont**, the CODYM (COnversational DYnamics Model) analysis uses simple behavioral state-based models (Markov Models) to capture the flow of information during different conversations, based on patterns in the lengths of alternating speaker turns. In addition to serving as a means for assessing and training healthcare providers, CODYMs could also be used to compare “conversational dynamics across language and culture, with the prospect of identifying universal similarities and unique ‘fingerprints’ of information flow,” said the study authors **Laurence Clarfeld** and **Robert Gramling**. Computing researchers and care professionals at the **University of Rochester** are **augmenting the difficult conversations around cancer treatment**. SOPHIE (Standardized Online Patient for

Healthcare Interaction Education) distilled best practices from 400 conversations between patients and oncologists to create an online virtual "patient" that helps physicians understand how to communicate effectively with late-stage cancer patients. Patients and their families frequently have to make difficult decisions using information that they fail to grasp. Care givers are under relentless time pressure. And these systems can help everyone.

## EARTH AND SPACE SCIENCES

Machine learning methods are increasingly being used in geo and earth sciences though these changes may not yet be reflected in teaching curricula. A **succinct round-up** by **Jacob Bortnik** and **Enrico Camporeale** of earth and geoscience papers using machine learning can help you identify gaps in your training (if you're worried you might have missed something) or help you write a syllabus.

The **use of ML in extreme heat events** got a test run recently when an international team of 27 scientists examined the links between human-induced climate change and the recent Northwest (US) heat wave through a rapid attribution study. (The study first had to overcome this problem: because the observed temperatures were far above what scientists had ever seen before, or even thought possible, the models initially couldn't replicate them.) Results showed the extreme temperatures – called "a mass casualty event" responsible for as many as 200 deaths in Washington and Oregon, 2000 deaths in Canada and massive fish and shellfish die-offs – were made at least 150 times more likely to occur because of climate change. "Basically without climate change, this event would not have happened," **Friederike Otto**, one of the study authors and a climate scientist at the **University of Oxford**. In related work, **Vivek Shandas** from **Portland State University** has been studying **the effect of heat islands on low-income and homeless populations**, and his data gathered during the recent heat wave showed a 25-degree differential in high temps between poorer and wealthier neighborhoods.

Earthquake prediction is one of the holy grails of applied geoscience so it's no surprise that **machine-learning models are being deployed** to hopefully, someday warn of impending doom ahead of the big ones. The recent Christchurch NZ earthquake allowed **Ellen Rathje**, an engineer at **University of Texas at Austin** and PI on **DesignSafe** to test her model. It correctly predicted the amount of lateral movement, which is a solid incremental step that proved the value of placing seismic sensors around the at-risk city and demonstrated the ML methodology required to process the immense volume of data.

There is an interesting wrinkle in the conversation about what happens when scientists are better able to predict expensive catastrophes: insurers have reason to raise rates or discontinue insurance altogether. This is already happening in the expensive, wild-fire stricken counties of **California wine country**. Wildfires can cost upwards of $13 billion per fire, most of which is born by insurers and reinsurers. In the research space, **SilviaTerra** **creates forest maps** that prioritize fire prevention according to vegetation type and location. **Chooch AI** detects the **first signs of fire** using computer vision and satellite imagery. **Squishy Robotics** created **a sensor ball** that can be dropped into a wildfire to take critical temperature readings. A couple other startups (**Zesty.ai** and **Kettle**) have started using deep learning technology and AI to evaluate wildfire risk (and provide reinsurance), but this is definitely a research space begging for partnerships with social scientists. The consequences of deploying more accurate models will cause insurers to raise rates or abandon markets, which will upend residential real estate (most mortgage lenders require property insurance), challenge the economic viability of risky areas, and leave some individuals and companies suddenly under water financially and/or climatologically.

## PROJECTS SMALL AND LARGE

**Taha Yasseri** and colleagues recently **published their analysis** of 150,000 anonymized eHarmony

UK dating app customers, mostly from London and all app-declared heterosexual (for purposes of the study). Successful matches, they found, are more likely when profiles have similar smoking habits and similar desire for children, and not as likely for match similarity between traditional status indicators like income or education. Most dating apps have more men than women male than female participants, which **runs counter to broader trends** where greater women's achievement has fostered an "increasing cohort of successful women are chasing a shrinking number of high-value, commitment-averse men," according to journalist and **University of Cambridge** Ph.D. student **Vincent Harinam**. The assortative selection in the dating game is more math than magic but, as the single among us know, dating is a *project*.

Building an IKEA bookshelf with a robot helper is an easier project with a narrow definition and specific tasks. Engineers at the **University of Southern California** crafted **just that machine**. **Sean Gallagher**, an editor at online magazine **Ars Technica**, is doing **something similar** but without the robot, crafting machine learning algorithms to help write headlines. Gallagher's **training set**: 5,500 headlines that have gone through Ars' A/B tests during the past five years.

When the project is larger and more abstract, like making urban transport more bike-centric and less car-centric, network analysis has a critical role. **Michael Szell** and colleagues at **IT University of Copenhagen** **simulated synthetic bicycle networks** that overlay streets in 62 international cities, finding that it took time (with initially decreasing returns) to grow to critical mass. Related, a team of Danish, Norwegian and American researchers led by **John Thogerson** from **Aarhus University** found that **pervasive infrastructure plays a role in habitual car commuting**. It's not enough to have a good plan, bike network supporters also need persistence.

Business-oriented network scientists **Marc Santolini**, **Christos Ellinas** and **Christos Nicolaides** **looked closely** at the inevitable delays that occur in large-scale engineering projects. The "perturbation cascades" of mishaps and delays causing more mishaps and delays are an opportunity for network-science frameworks to help manage large, time-sensitive, expensive projects. And maybe, in time, to help manage smaller projects too.

## ASTROPHYSICS, QUANTUM PHYSICS AND EFFECTIVE COMPUTATION

An astronomical collaboration using open data, the U.S. National Lab system (**Argonne**, **Oak Ridge**), universities (**University of Chicago**, the **University of Illinois at Urbana-Champaign**) and industry partners (**NVIDIA**, **IBM**) **showed that** a workflow using "Data and Learning Hub for Science, a repository for publishing AI models, with the Hardware-Accelerated Learning (HAL) cluster, using funcX as a universal distributed computing service" accurately detected the gravitational wave patterns associated with all four black hole mergers observed in one month's worth of LIGO data with no false positives. Processing took less than 7 minutes with 64 NVIDIA V100 GPUs sharing the computational load. One of the explicit goals of projects like this is to make this type of data, modeling, and computational infrastructure readily available to researchers. Shout out to **Daniel Katz**, one of the **paper's** authors and long-time DSCN reader.

The work these teams have done to build partnerships, infrastructure, and workflows has the potential to supercharge *any* research for which the amount of available data is (far) greater than the perceived available computing power.

Another recent **Harvard**/**Cornell** paper out of the quantum physics world that could have broader implications developed **"Correlator Convolutional Neural Networks (CCNNs)**" that are "a set of nonlinearities for use in a neural network architecture that" reveal "directly interpretable" features that can be linked to physical observables. In other words, neural networks can become accurate at the tasks they are put to, but with highly complex data (which is usually an attribute of the tasks to which they are put), it's hard to know which attributes of the underlying data are driving predictions, especially if the training data are simulated. When the attributes that drive predictions are unknown or unknowable, it's difficult to derive scientifically sound reasoning based on the CNNs performance. This is a problem not only in physics, but in any scientific application of convolutional neural net approaches. Though **this paper** is specifically applicable to quantum physics, the motivating impetus is much more broadly shared and is worth a skim.

The environmental impact of computationally intense modeling is substantial. Training a certain NLP model can release more CO2 than driving a car...for its entire lifetime. **Ryan Hamerly** of **NTT Research**, currently visiting at **MIT's Quantum Photonics Laboratory**, thinks that the future of computationally intensive neural net processing may be in optical computing, which would mean the past (fiber optics) could be new again though this time for computing, not communicating. There are still major hurdles optical computing has to overcome, like size (there's no nano in optical computing) and noise. It's a **fascinating explanation** of why optical computing is particularly well-paired to linear algebra, written so well that you will understand the concepts even if you never took linear algebra. The upshot is that hardware innovation will continue to be an important aspect of the data science ecosystem. Neither GPUs nor TPUs nor any other processor should be considered the agreed-upon go-to for applied AI.

## FOLLOW THE MONEY

**$702.4 million** Research-Specific External Funding Sources -> **University of Iowa**, a 31% yearly increase despite, or maybe because of, the global pandemic. **Iowa State University** received $231.1 million, a 13% yearly increase.

**$300 million** **Helen Diller Family Foundation** -> **University of California-Berkeley** for a new 772 bed “Anchor House” dorm for transfer students that critics say is “too big and too fancy”

**$117.2 million** **University of Illinois-Chicago** -> **LMN Architects**, **Booth Hansen** will break ground on UIC's 135,000 square foot Computer Design Research and Learning Center. The building sits next to Memorial Grove on the school's historic campus designed by **Walter Netsch**.

**$40 million** **National Science Foundation** -> **University of Connecticut**, **University of Wisconsin**, **University of Georgia** to jointly establish the **Network for Advanced Nuclear Magnetic Resonance**.

**$23.2 million** **Foundation for Food & Agriculture Research**, **Nestle**, **Dairy Management Inc.**, **Newtrient** -> **Soil Health Institute**, **Cornell University**, **University of California-Davis**, **University of Texas A&M AgriLife Research**, **University of Wisconsin-Madison**, **University of Wisconsin-Platteville**, **University of Vermont**, **USDA Agricultural**

**Research Service** and **Northwest Irrigation and Soils Research** for six year funding for the Net Zero Initiative, an industry-wide effort to adopt practices and technologies that reduce greenhouse-gas emissions and improve environmental health.

**$17 million** **Bloomberg Philanthropies** -> **San Francisco**, **Bogotá**, **Amsterdam**, **Mexico City**, **Reykjavík** and **Washington DC** city governments to establish "innovation teams" that work across agencies and use technology to improve government services.

**$3 million** **National Science Foundation** -> **Center to Stream Health In Place** for research that would allow care clinicians to monitor patients using wearables.

**$2 million** **Fellows Fund**, **Silicon Valley Future Capital**, **Eastlink Capital**, plus individual Silicon Valley executives -> **InsightFinder**, an IT services startup that uses machine learning to prevent service outages, founded by **North Carolina State University** associate professor **Helen Gu**.

**$1 million** **National Science Foundation** -> **Georgetown University Cyber SMART Research Center** was selected as an Industry University Cooperative Research Center (IUCRC). Established in 2019, the center's name reflects its multidisciplinary approach and the five core elements covered in all its research: science, management, application, regulation and training (SMART). Cyber SMART will open a second site at the **University of Notre Dame**.

**$120,000 per year** **University of Colorado** -> Ralphie the Buffalo Mascot. According to reporter **Emily Caron**, the spectacle of colleges' live animal mascots has gotten expensive.

**$24** Kansas residents age 60+ -> **Wichita State University** will start charging local senior citizens to audit courses. Fees range from $8-68 per credit hour and most courses are 3-credit.

**$0** **University of Michigan** and development partner **Bedrock** -> **Detroit** after the developer canceled a $300 million project to jointly build the Detroit Center for Innovation at a former jail site downtown. The developers will look to develop the Center at another site in the city.

## NEW PROGRAMS

**Artificial Intelligence degree** @ **Wayne Community College**.

**University of Connecticut** -> adds new undergraduate major in Robotics Engineering, open to students Fall 2022.

**AI-on-5G** -> new lab from **Google Cloud** and **NVIDIA**.

**6G@UT** -> new research center at **University of Texas at Austin** with **Samsung**, **AT&T**, **NVIDIA**, **Qualcomm** and **InterDigital**.

**Unnamed digital storytelling R&D group** -> **Netflix**, one of **Paul Debevec's** responsibilities as the newly hired director of research. Debevec will remain an adjunct professor at the **University of Southern California Institute for Creative Technologies**.

**Center for Research toward Advancing Financial Technologies (CRAFT)** -> **Stevens Institute of Technology** and **Rensselaer Polytechnic Institute** create fintech research lab

with **NSF** funding.

**Unnamed digital storytelling R&D group** -> **Netflix**, part of **Paul Debevec's** responsibilities as the newly hired director of research. Debevec will remain an adjunct professor at the **University of Southern California Institute for Creative Technologies**.

**AI in Medicine self-paced online certificate** -> **University of Illinois at Urbana-Champaign**, for healthcare professionals, costs $750.

**Data Science and Analytics Minor** -> **University of Mount Union**.

**Data Science for Environmental Applications certificate** -> **Western Washington University** 9-month asynchronous, remote, ~$6508.

**Undergraduate Data Science Major** -> **University of Illinois at Urbana-Champaign** is one step closer to a new Data Science degree program with approval from the University Senate. Additional approvals are required from the **Illinois Board of Higher Education** and from the **University of Illinois System** before a Data Science degree can be part of the curriculum. The earliest students can expect the degree coursework is fall of 2022. The Data Science degree will be offered as part of a double major degree, rather than its own program in order to "allow students to have the opportunity to further their studies through a data science lens."

**University wide *major* restructuring** -> **University of Wyoming** faculty Senate has proposed a major restructuring that would create **a School of Computing**, a Center for Entrepreneurship and Innovation, and a Wyoming Outdoor Recreation, Tourism and Hospitality Initiative while eliminating 75 faculty and staff positions, including 10 department heads. Will save $13 million annually.

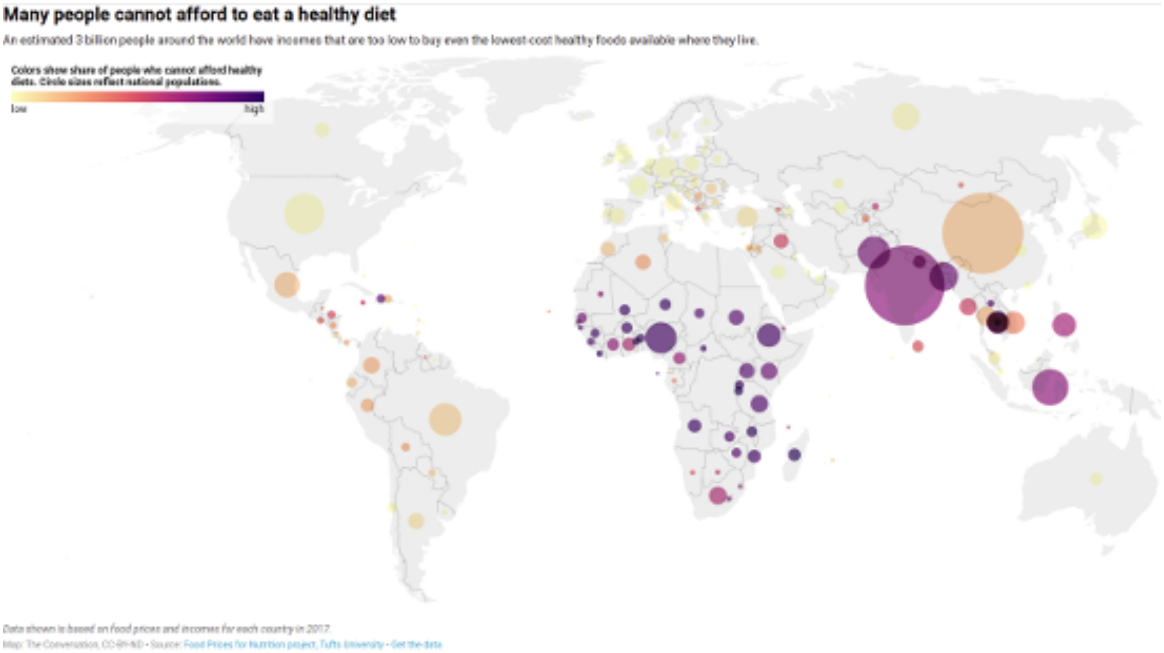
**Integrative and Quantitative Biosciences Accelerated Training Environment (InQuBATE) Predoctoral Training Program** -> **Georgia Institute of Technology** has an **National Institutes of Health** grant to help transform the study of quantitative- and data-intensive biosciences at the school.

**Center for Quantum Technologies** -> **Purdue University** with in-state academic partners: **Indiana University Bloomington**, the **University of Notre Dame**, and **Indiana University Purdue University-Indianapolis** (IUPUI).

**FUTURE-MINDS-QB** -> **University of Illinois at Urbana-Champaign Carl R. Woese Institute for Genomic Biology** established a bridge program streamlining a path from a master's degree at **Fisk University**, a historically Black university in Nashville, to a doctoral degree at University of Illinois. The program is currently accepting applications.

## DATA VISUALIZATION OF THE WEEK

The Conversation, William A. Masters and Anna Herforth from July 9, 2021




## Events

See the ADSA Events Page for more details and more opportunities.

**2021 SIAM Block Community Lecture Presented by Jonathan Christopher Mattingly**
**Online** July 20, starting at 1:30 p.m. Eastern. The talk is titled "Can You Hear the Will of the People in the Vote? Assessing Fairness in Redistricting via Monte Carlo Sampling." [free, registration required]

**United Nations 2nd Open Science Conference 2021**
**Online** July 21-23. "Policy makers, main IGO actors, librarians, publishers and research practitioners will engage into a public dialogue focusing on what Open Science has learned from COVID-19 and how this can be applied into actions addressing the global climate crisis." [registration required]

**South Big Data Innovation Hub All-Hands Meeting**
**Online** July 28-30. "Conference tracks align with the priority areas of the **South Big Data Hub** and are each led by a game changer within their area of expertise. These leaders are building sessions to spark conversations and create collaborative opportunities." [registration required]

## Deadlines

**Contests/Award**
**$100M Prize For Carbon Removal**
"XPRIZE Carbon Removal is aimed at tackling the biggest threat facing humanity - fighting climate change and rebalancing Earth's carbon cycle. Funded by Elon Musk and the Musk Foundation, this $100M competition is the largest incentive prize in history, an extraordinary milestone." Deadline for submissions is October 1.

**Studies/Surveys**
**NIST Proposes Approach for Reducing Risk of Bias in Artificial Intelligence**

"In an effort to counter the often pernicious effect of biases in artificial intelligence (AI) that can damage people's lives and public trust in AI, the **National Institute of Standards and Technology** (NIST) is advancing an approach for identifying and managing these biases — and is requesting the public's help in improving it."

**RFPs**

**National Endowment for the Humanities Humanities Connection program**
"Awards support innovative curricular approaches that foster partnerships among humanities faculty and their counterparts in the social and natural sciences and in pre-service or professional programs." Deadline for applications is September 14.

**Conferences**

**UIST Student Innovation Contest**
**Online** "We explore how novel input, interaction, actuation, and output technologies can augment interactive experiences! This year, in partnership with Sony Interactive Entertainment, we are seeking students who will push the boundaries of input and output techniques with the TOIO micro robot platform." Deadline to apply is August 2.

## Tools & Resources

**Good and Bad Monitoring. You remember that time you…**
*Eric Pyle* from July 2, 2021
"Take a look at your project's compilation warnings. If you're using NPM, you'll see the impossible to resolve deprecation warnings a mile long and quickly realize how much people ignore issues. Still, something has everyone convinced that people actually want to fix things. What leads to this massive disconnect? Bad monitoring. Let's go over traits of a good and bad system."
**We now list 198 fellowships, including link, description, amount, eligibility criteria, deadline.**
*Twitter, Denis Wirtz* from July 9, 2021
"We have updated and expanded our database of PhD fellowships."

## Careers

See the ADSA Jobs Page for more opportunities.

**About Us**: The Data Science Community Newsletter was founded in 2015 in the Moore-Sloan Data Science Environment at NYU's Center for Data Science. We continue to be supported by the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation through the Academic Data Science Alliance. The newsletter is written and the content is compiled by the Academic Data Science Alliance. Our archive of newsletters is at cds.nyu.edu/newsletter and is the process of transitioning to another, permanent location. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.