ADSA Data Science Community Newsletter

Data Science Community Newsletter features journalism, research papers and tools/software for February 24, 2022.

Please let us (Micaela Parker, Catherine Cramer, Brad Stenger, Laura Norén) 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.

ADSA VIRTUAL SESSION **Creating Inter-Institutional Data Science Pathways: Streamlining Access to Data Science Education in California**

A panel discussion, held last December 15, on how the California public college and university system — with 3 million students enrolled — is figuring out how to offer data science equitably across a diverse range of institutions and communities.

Write-up
Full-session Youtube video

ADSA SPRING MEETING

For those of you who are coming to **ADSA's spring meeting** on March 7-9, see you soon!

Brad and Laura will not be in Irvine, but Micaela will be. She is looking forward to getting feedback about the DSCN in person.

M | MIDAS
MICHIGAN INSTITUTE FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN
2022 Data Science Fellows Program
Applications Open!

This 2-year program provides outstanding early career researchers with intensive data science experience as they prepare for independent research and faculty positions. Apply now!

WHO'S HIRING? [SURVEY]

Last week, **Avi Kak**, professor of Electrical Engineering and Computer Science at **Purdue**, got us thinking about whether liberal arts universities are turning into "**glorified approximations of trade schools**" by hiring in technical departments quickly while the arts and humanities struggle to support their graduate programs.

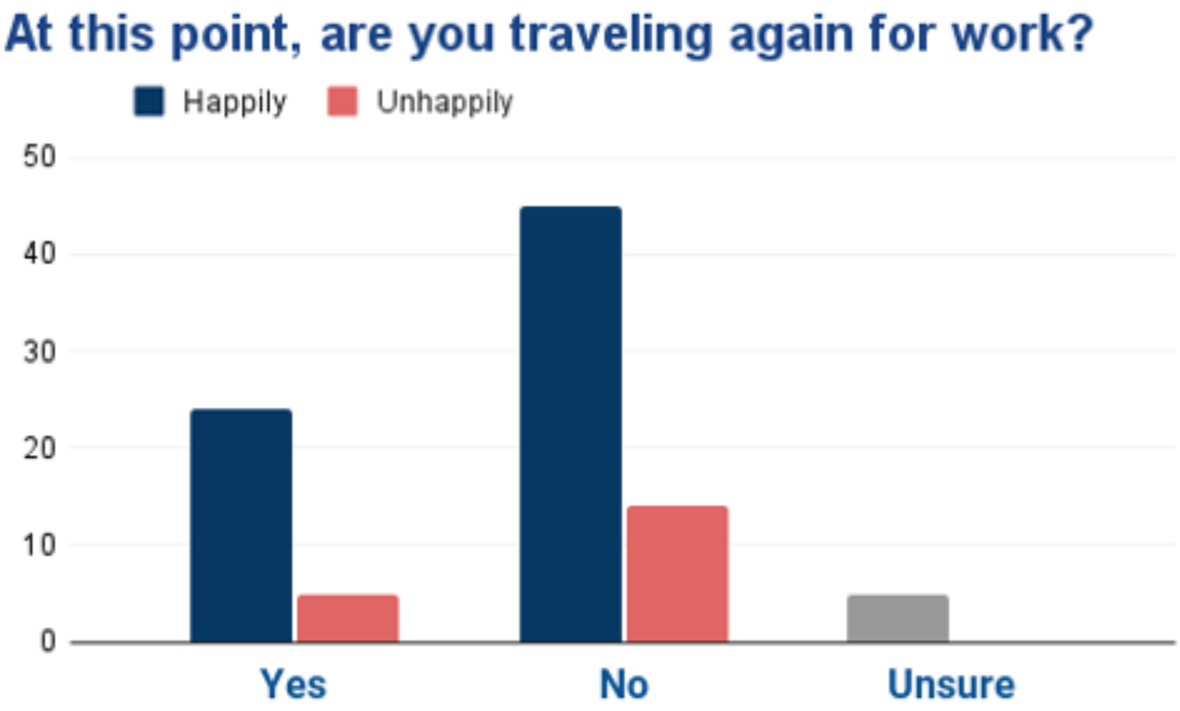
Help us quantify Kak's hypothesis. **Take Our Survey**

BUSINESS TRAVEL MAY NOT BE THE SAME EITHER [SURVEY RESULTS]

The results of our first three surveys suggest a new campus normal characterized by less travel, more focus time, and work-life balance.

The most impactful change is to the weekly rhythm. Faculty and staff will noticeably reduce the number of days they spend on campus, mirroring a national trend in more WFH time. The modal category of days per week spent on campus is only 1-2 days post-pandemic in a group that spent 3-5 days per week on campus pre-pandemic, but travel to conferences and workshops will also change for *some*.

Travel expectations were sharply mixed. The modal category noted that they aren’t back to traveling for work yet and are happy about it. Combine this with the small number of people who have started back on the road but are unhappy, and it looks like 58% of you don’t want to travel for work. But…a healthy 42% of you do want to be back on the conference circuit. This puts event organizers in the tricky position of having to offer high quality in-person events and high quality remote experiences. I predict the big disciplinary conferences and events like NIPS, ICLR, and ICML will stay big and stream content, but smaller conferences and workshops may be forced to get leaner and scruffier.



When combined with the moral imperative to reduce air and car travel in order to save the planet, expect vocal campus groups to use evidence of lack of interest in travel and 5-day-a-week commuting to strongly oppose daily commutes and in-person conferences as we establish the new academic expectations. We do need to normalize putting climate impacts right up in the top three considerations for any decision.

FUN FROM THE ARXIV — THREE ERAS OF ML

There's a fun, **highly teachable new paper** on arXiv this week that shows three distinct eras of ML based on the compute power required to train models: "Since the advent of Deep Learning in the early 2010s, the scaling of training compute has accelerated, doubling approximately every 6 months. In late 2015, a new trend emerged as firms developed large-scale ML models with 10 to 100-fold larger requirements in training compute. Based on these observations we split the history of compute in ML into three eras: the Pre Deep Learning Era, the Deep Learning Era and the Large-Scale Era." (h/t **Twitter, Haydn Belfield and Lennart Heim**

Featured Job

جامعة نيويورك أبوظبي
NYU ABU DHABI

Postdoc - 2 year. NYU Abu Dhabi, Social Research and Public Policy (health). Abu Dhabi, UAE (no remote option).

OSTP: FROM ERIC LANDER TO ALONDRA NELSON AND FRANCIS COLLINS

Two weeks ago we reported that **Eric Lander** had resigned from his leadership position at the **White House Office of Science and Technology and Policy** (OSTP) over Lander's callous, demeaning leadership style that violated the Biden White House's stated guidelines.

Now the **Biden Administration** announced that **Alondra Nelson** will lead the OSTP and **Francis Collins** will serve as Science Advisor to the President and Co-Chair of the **President's Council of Advisors on Science and Technology** (PCAST) "until permanent leadership is nominated and confirmed". Nelson is a sociologist and Collins is a physician who just stepped down as head of the **National Institutes of Health** in December 2021. This joint nomination with Nelson in an interim role deeply upset **H. Holden Thorp**, the editor of **Science** who wrote an op-ed to express that Nelson "represents the future of American science" and should have been nominated to leadership without the training wheels of "interim" or co-lead to slow her down.

Elsewhere in the **White House** science policy machinery, the gears continue to turn on the new expectations researchers will have to meet when disclosing their conflicts of interest with foreign countries, especially China, Russia, and Iran. Any salary or honorarium must be disclosed, but even receiving a commemorative plaque could trigger scrutiny and possibly jeopardize a researcher's ability to receive grants from the U.S. government. The global nature of the scientific community is not well paired with national funding and data control regimes.

Featured Job

ALLEN INSTITUTE

Neuroscience needs data scientists. Fellowship at the Interface of Data & Neuroscience (Scientist I position - Allen Institute, Seattle).

UC BERKELEY ENROLLMENT DISPUTE

As we foreshadowed in a previous newsletter, **University of California, Berkeley's** new student housing project raised flags with Alameda County's environmental review process. The ongoing dispute could force the university to hold back 3050 slots for first-year and transfer students.

Town/gown politics are generally kept quieter.

THREATS TO TENURE AT STATE SCHOOLS MOUNT

Former Secretary of Agriculture under former **President Trump** and climate change skeptic **Sonny Perdue** has been nominated to serve as chancellor of the **University of Georgia** system. He will likely take the position later this week despite strong opposition from student groups, UGA faculty, and the **American Association of University Professors** (AAUP). A **Mother Jones** article summarizes Perdue as, "an agribusiness tycoon … best known for his political work as a former two-term Georgia governor during which he urged pay cuts for school teachers; expressed nostalgia for the Confederacy; pioneered voter-suppression laws; responded to a brutal drought by holding a public ceremony to 'pray for rain'; and promoted policies that made life miserable for immigrants."

Outgoing **University of Wisconsin** chancellor **Rebecca Blank** used her farewell address to warn regents that the state's political polarization is the “greatest existential threat" to the university system. During her tenure **Governor Scott Walker** slashed funding and moved to eliminate tenure. The regents reportedly gave her a standing ovation.

Eric Hartzell, President of **University of Texas at Austin**, was compelled to defend the practice of tenure after **Dan Patrick**, the Lieutenant Governor for the **State of Texas**, proposed ending it at state universities.

Featured Job

See the ADSA Jobs Page for more opportunities.

Admin. Associate Director. University of Wisconsin-Madison American Family Insurance Data Science Institute. Madison, WI (Some remote work possible).

ANDREW NG — TACTICALLY PATCH LARGE ML MODELS

Andrew Ng, founder of **Google Brain** and former head of data science at **Baidu**, talked about how to improve extremely large AI models. He recommends both 1) identifying which elements of the training set are generating the most noise and 2) adding small-ish (N=100) targeted training set patches to improve performance in human-identified edge cases. Tactical patching isn't new, but it's coming up again because it is still necessary even in really large scale models. It's important to make sure ML models are keeping up with our best intentions as humans. Of course, in order for extremely large ML models to deliver results that work equally well across diverse populations in inclusive ways, the models' trainers' need to know how to identify those types of diversity and carefully tune the model by feeding it new data.

MINNESOTA WILL WARM THE MOST

In a new study by researchers from the **University of Minnesota** and **University of Alabama**, scientists found that within the next 80 years, central Minnesota will have 55 fewer days of snow cover, will warm by 11 degrees in the winter, and jump 7 degrees in the summer. Its reputation for being too cold could turn into a reputation for being one of the most appealing locations to buy a home, especially if insurance companies stop renewing policies for properties in wildfire and flood-prone areas.

COVID DATA PROBLEM

As we take a deep breath and look out at the long term COVID horizon, there are some big data problems:

1) **Trevor Bedford**, an epidemiologist at **Fred Hutchinson Cancer Research Center** estimates that "only about 20% to 25% of omicron infections in the US get reported," largely due to people opting for home testing and no-testing.
2) **The New York Times** raised questions about how much data the **U.S. Centers for Disease Control and Prevention** (CDC) and local departments of health aren't releasing to the public or to the broader scientific community, especially about wastewater data that can be a better real-time source of local prevalence levels than testing data.
3) There is still much to be desired in terms of tracking the emergence of new genetic variants in the US and globally. Starting out at a scattershot ~1% in the early pandemic, the US was testing 5-10% of samples by November 2021.
4) **Eduoard Mathieu** from **Our World in Data** wrote a World View opinion column in **Nature** saying that the international governments' medical oversight, led by the **World Health Organization** (WHO), should focus efforts on data collection, not on dashboard design.

If we have any hope of instituting a reasonable dial-up/dial-back system that is responsive to threatening new variants, we need more scientists to have access to more data in as close to real-time as possible.

CAN AI MAKE STORE-BOUGHT TOMATOES TASTE GOOD?

Full disclosure — I am a tomato snob. No matter how delicious store-bought tomatoes look, they never compare to the way home grown tomatoes taste. This is largely the fault of supply chain logistics — tomatoes are sensitive to bruising, mold, cold temps, slugs, birds, various insects, too much watering, and getting poked or scratched by anything at all. Most selective breeding has produced tomatoes that survive this supply chain looking red and ripe … though they still have lackluster taste. Now **Marcio Resende** and **Harry Klee** are developing AI to select tomato strains that taste good *and* survive the supply chain. In this case, AI generates predictions that streamline the development cycle.

WHILE YOU'RE IN THE STORE, MASK UP. CLEARVIEW AI MAY ALREADY HAVE YOUR FACE IN ITS DB.

Clearview AI, one of the worst actors in the AI space, is apparently using a pitch deck that explains the company has built an architecture capable of cataloging and writing ML models to identify 100 billion faces. That means everyone on Earth. Clearview hopes to sell access to this data to retailers and gig economy employers/platforms. This is yet another reason to continue masking in stores.

THE UK NHS THINKING ABOUT ALGORITHMIC IMPACT ASSESSMENTS; 80% OF AMERICANS ARE UNWITTINGLY REPRESENTED IN A DATASET OWNED BY A PRIVATE EQUITY FIRM

The **Ada Lovelace Institute** has delivered a detailed proposal for implementations of Algorithmic Impact Assessments, purportedly for those who want to use **National Health System** data in the UK. The template they propose is fairly flexible and could be the starting point for algorithmic impact assessments in any field, not just healthcare and not just in the UK. Having recently been more involved in local government "community outreach" sessions that sound a lot like the "participatory" sessions outlined in the proposal, I would like to see more accountability tied to collecting data and being responsive to it. Too often, these sessions are seen as a check-box exercise where real grievances and concerns are raised, but there's no mechanism for sorting out which grievances are reasonable, which must be addressed in full, which can be addressed partially or not at all, and how to sort out differences between stakeholders.

For instance, let's say I have a COVID vaccine that will outright prevent individuals from ever getting

COVID, but in order for them to work right, I need to run a predictive model that is extremely energy inefficient — worse than bitcoin mining. The AIA proposed here will reveal that environmental cost, which is pretty cool. But it's unclear which mitigations would be considered reasonable. Any process that revolves around identifying and mitigating disparate impact along multiple vectors is going to be extremely challenging. This does not mean we shouldn't face these differences head on. But it does mean that algorithmic impact assessments require collaborative, non-colonialist politics, and a willingness to make hard calls.

This is a much healthier conversation than the one we largely aren't having around the largest health database in the US. Pun intended.

A company called **MarketScan** holds health data on 80% of Americans and has recently been sold by the defunded **IBM Watson** project to private equity firm **Francisco Partners**. The health data are reportedly "anonymous." "Anonymous" is in quotes not because we believe patients' names are still in the records — they probably aren't — but because it is difficult to fully anonymize data of this sort *and* because individual-level health data is sensitive even when it is truly anonymous. **WBUR** in Boston **interviewed** **Casey Ross**, who **wrote** a pay-walled article on *STAT* about how MarketScan has quietly become a power broker of health data. **Ernie Ludy**, who founded MarketScan as a “sacred trust" thinks patients should, at a minimum, be notified about and compensated for the value of their health records. There is a huge difference in housing data within the NHS, a government agency dedicated to the well-being of citizens, and having a private equity firm figure out how to squeeze every last penny out of it.

DEEPMIND WORKING ON NUCLEAR FUSION

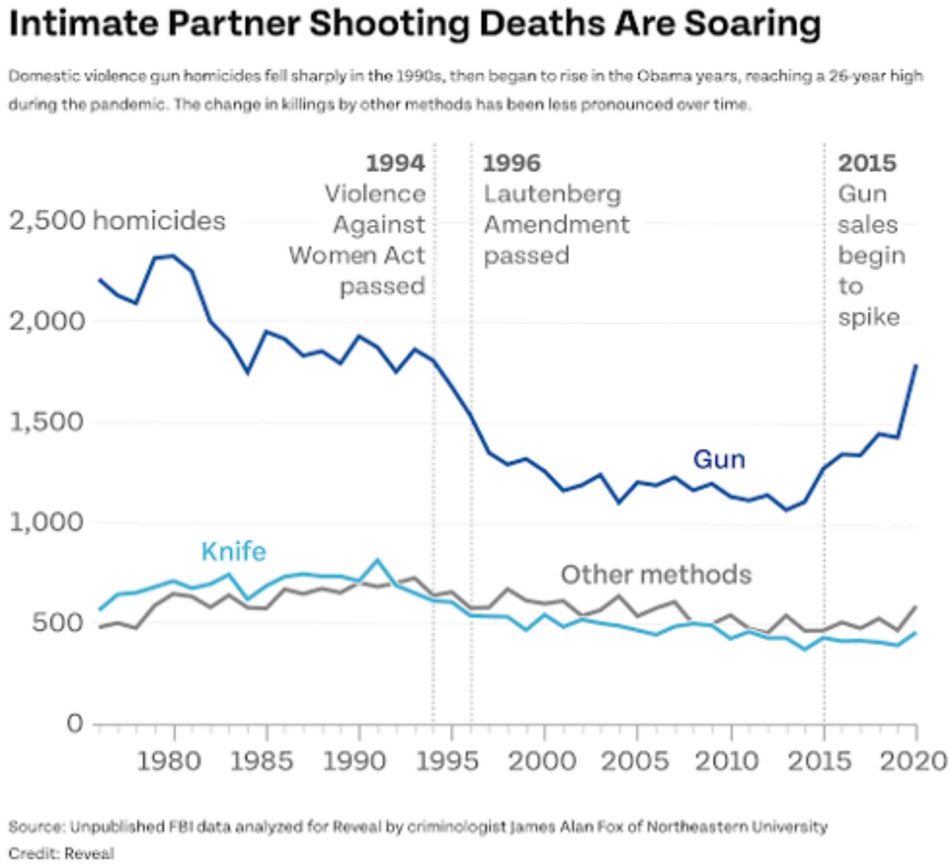
In partnership with nuclear physicists at **EPFL** in Switzerland, **DeepMind** has been working on **reinforcement learning algorithms** to identify the right magnetic fields to confine plasma at temperatures hotter than the sun's core to facilitate fusion. EPFL has a donut-shaped chamber designed to facilitate fusion research called a tokamak, one of only a handful in the world. Like so many sub-fields in physics, time spent with the tool is rare and precious.

NEW PROGRAMS, FOLLOW THE MONEY

Click through to access **a structured spreadsheet** of New Programs and money moving around in academic data science.

DATA VISUALIZATION OF THE WEEK

Reveal, Jennifer Gollan from October 26, 2021



Intimate Partner Shooting Deaths Are Soaring

Domestic violence gun homicides fell sharply in the 1990s, then began to rise in the Obama years, reaching a 26-year high during the pandemic. The change in killings by other methods has been less pronounced over time.

Source: Unpublished FBI data analyzed for Reveal by criminologist James Alan Fox of Northeastern University

Credit: Reveal

Deadlines

Conferences
[It took forever, but applications are finally open for the "Astronomical Software Development" workshop that we're hosting at @FlatironCCA, May 16–20, 2022!]
"Check out the website for the link to the application form & all the other details: https://code.astrodata.nyc" Deadline for applications is February 28.

Education Opportunities
Applications for Bloomberg's #DataScience Ph.D. #Fellowship for the 2022-2023 academic year are now being accepted
"submission deadline: April 15"
Contests/Award
$10K prize to winner of native birdsong coding competition
"Hawaiʻi has lost 68% of its bird species. A bioacoustics laboratory at the **University of Hawaiʻi at Hilo** that specializes in the ecology and conservation of Native Hawaiian forests and birds is co-sponsoring a competition to help develop song detection algorithms for Hawaiian birds." Deadline for entries is May 17.

Tools & Resources

Tying Artificial intelligence and web scraping together [Q&A]
Beta News, Wayne Williams from February 16, 2022

"While web scraping has been around for some time, AI/ML implementations have appeared in the line of sight of providers only recently. **Aleksandras Sulzenko**, Product Owner at **Oxylabs.io**, who has been working with these solutions for several years, shares his insights on the importance of artificial intelligence, machine learning, and web scraping."

Announcing AI Blueprints Public Preview

cnvrg.io Blog, Yochay Ettun from February 21, 2022

AI Blueprints are open-source pre-assembled machine learning pipelines that are built by **cnvrg.io's** ML experts, but they can also be created for custom uses by any user. They're for developers who want to build machine learning-powered apps and services for popular business uses like recommender systems, sentiment analysis, object detection, and many others."

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. Our archive of newsletters is at https://academicdatascience.org/resources/newsletter. Our mailing address is 1037 NE 65th St #316; Seattle, WA 98115.