Data Science Community Newsletter

16 April 2020 **Data Science Community Newsletter**

Please let us (Laura Noren, 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 and to NYU's Center for Data Science.

Data Science News

University Data Science News

A **new pre-print** by **Kim Weeden** at **Cornell University** and **Benjamin Cornwell** used a network simulation to understand **the implications of sending students back to campus in the fall**. Assuming the US does not see a massive new wave of infections, most people are likely to remain naive to the virus in September/October 2020. So should we send students back? Using transcript data from Cornell's students, Weeden and Cornwell discovered that even if we only consider interactions in class, there's no easy way to prevent COVID-19 from consuming campuses. Campus networks are highly clustered by major, but still: "the average student can 'reach' only about 4% of other students by virtue of sharing a course together, but 87% of students can reach each other in two steps, via a shared classmate. By three steps, it’s 98%." Even if classes larger than 100 students went online, transmission would slow, but would still eventually hit everyone. Some have noted that the study isn't representative of commuter schools, but keep in mind that Weeden and Cornwell only looked at transcript data, so it is likely to be representative of students who take a full load of courses, whether they live on campus or not.

As a previous title of this newsletter suggested, **Harvard** researchers have now produced simulations suggesting that the curve will oscillate, probably into 2022. Because social distancing may have kept much of the population from being exposed, almost all of us are still susceptible. With Germany and Spain looking to start re-opening parts of their economies that have been locked down and the US not far behind, we will see greater chances for people to contract COVID. At some times, in some places, new flare ups will be so serious they require lock downs to quell them. The new simulation suggests that **waves of high caseloads could persist into 2022** and that surveillance measures should remain in place until at least 2024. The coronavirus will mutate and immunity could be limited to a 2-3 year window, after which something like SARS-CoV-3 could emerge and start another kind of infection cycle. Alternatively, we could remain immune to SARS-CoV-X and never face a newly mutated coronavirus capable of making us ill. Simulations like this one are begging for better serological testing in order to figure out the true R0 for SARS-CoV-2.

Boston University is **looking at a Fall 2020 return-to-campus date** though can pivot to January 2021 if necessary. Most schools have not announced their target restart dates.

The **University of Oregon** has announced that it will **layoff 282 employees**. Senior administrators will take a salary cut. The University generates 25% of its revenue from housing and sports, both of which have been canceled.

A loose 'do-acracy' of researchers are compiling coronavirus data and projects at **Science Responds**. They are a little light on social science and political science coronavirus projects, so if you happen to be interested in working on that angle, I'm sure they would be happy to let you join their scientific flock. As an full-throated supporter of radical interdisciplinarity in data science, I think there's mutual benefit involved in adding more social science research.

Nature will be moving to **participate in Open Access (OA) publishing in January 2021** for scientists who want it or for funders who have demanded it. This follows pressure from an ambitious OA initiative called Plan S that has been pushing to make all scientific publications available to the public upon publication.

Professor **Akiko Iwasaki** of **Yale University** opened her DMs to graduate students and postdocs struggling with their Principal Investigators (PIs). She was inundated and has **written up her qualitative findings** in a news article for *Nature*. The most alarming — but well-known issue — is that foreign graduate students and postdocs are manipulated more often and more effectively than their American-born counterparts. Iwasaki is concerned about the abject power dynamics between PIs and trainees: "PIs can hold trainees hostage through letters of recommendation and publications. International trainees are even more vulnerable, because PIs can hold them hostage with their visas." She recommends that department chairs implement changes in order to reward good mentors and withhold promotions from toxic PIs.

Adding to this conversation about the organizational sociology of academia, a **Stanford University** **study of three decades of US PhD dissertations** found that, "demographically underrepresented students innovate at higher rates than majority students, but their novel contributions are discounted and less likely to earn them academic positions." Iwasaki explained why extreme power imabalances are bad for scientific trainees. The Stanford team helps explain why invisible (to some) prejudices are bad for science. Always remember that we're in the scientific community. The work we do is bigger than ourselves or our own careers, but if we can't have careers, we can't contribute. Keeping the field open to as many people as are able to productively contribute should be understood as part of the scientific method.

Christopher M. Petrilli of **NYU** was the lead author on a study that found **age and obesity** are the two most serious comorbidities with COVID-19.

Company Data Science News

Google and **Apple** have **declared their intent** to offer contact tracing via their mobile phones. The idea goes like this: most Americans have either an Android phone or an iPhone and the phones are constantly sending location data back to Apple and Google. If people opted-in to share their positive test results, every phone owner they'd been near within the incubation window could be warned.

Those who have had a long exposure to a COVID+ person could even be advised to quarantine for 14 days. There are big questions about how marginalized communities would benefit — people in prisons, nursing homes, homeless shelters, and overcrowded housing conditions have proven to be at greater risk. I do not know exact percentages, but some of those populations undoubtedly do not have smart phones. There are even bigger questions associated with opting into a massive surveillance project.

Discussion about the **Apple** + **Google** plan to embed contact tracing capacity in their hardware **has been robust** (**more**). Commenters have mostly touched on privacy, civil liberties and massive surveillance. The privacy concerns are not so much that Google and Apple would get access to more of your personal data. That ship has sailed. But forcing people to allow themselves to be tracked in this way or even implying that they are insufficiently moral if they object is dangerous. First, it's not clear that this will actually work. Participants would have to opt-in. This is America. Many won't, or can't. Without large numbers the strategy won't be sufficiently effective. The people who choose not to opt-in are quite rightfully concerned that massive surveillance regimes should not be implemented on the fly without proper democratic debate. They have a point. My hot take is that all of the stress has us wrapped up in a rerun of a classic American genre: the technological fantasy in which we invent our way out of catastrophe...with an app. Generally speaking, that doesn't work. And when it does sort of work, it tends to work best for the wealthy. That seems likely here, too. Only people with smartphones can participate. Further, people who are able to largely self-isolate due to jobs outside the service sector may receive fewer false alarms. Those who are forced to move around more or work in grocery stores and pharmacies — delivery people, etc — will be likely to get more alarms. Will this make them healthier? Maybe...but only if it doesn't generate a series of false alarms that gets them out seeking tests that don't exist, or waiting for results that come back negative. Living through either of those scenarios would seriously mess with anyone's mental health. I could go on, but I'd like to politely suggest that we move the conversation past the privacy issue and onto all the other ways this is not a sufficient solution.

In existing projects involving massive surveillance that did not involve an opt-in consenting procedure (nor is there an opt-out opportunity), **Palantir** **has been working** with the **U.S. Centers for Disease Controle** (CDC) to evaluate health care needs during the pandemic. The company's revenue is **projected to hit $1 billion** this year, up 35% from 2019.

Reid Health in Indiana has become so **disillusioned with the accuracy of COVID-19 test results** that they have started reporting all the patients in containment in addition to those who have tested positive. They believe the tests have about a 30% false positive rate.

It's hard to blame **Amazon** for deciding **to try to develop testing capacity on their own**. They'd like to make sure they can provide timely tests for the 100,000+ Amazon employees.

Government Data Science News

If you would like to **participate in a serological study** to see if you were exposed to COVID-19, the **U.S National Institutes of Health** (NIH) has kicked one off. People who have *had* a positive

COVID test and those who are currently experiencing symptoms are not eligible. Everyone other American adult is. You'll have to **be comfortable dealing with your own blood**.

Iceland has tested about 10% of its population and found that **about 50% of the people who test positive are asymptomatic**. That's great news for the people who are essentially unaffected by the virus, nobody likes to be sick, but it's terrible for the rest of us. If people don't know they are contagious, they're less likely to take precautions like self-isolation and are likely to spread virus more widely because they aren't sick enough to realize that they need to be extra careful.
A **study** published in the *British Medical Journal* with new Chinese data found that an even greater percentage (78%) may show no symptoms. More serological studies cannot come fast enough.

In **Germany**, the COVID-19 caseload is low, likely due to their Korea-like testing schema. They are conducting **500,000 tests per week and utilized intensive contact-tracing early**. Their contact tracing does not rely on mobile phone surveillance yet, but that is part of future plans, though many are concerned about the Stasi-like similarities. German health officials relied on working with those who tested positive to figure out where they had been, who they had talked to, and, occasionally, they even compared virus genomes to make sure they knew who got what from whom. Their case rate is low enough that Merkel is planning to unlock the country slowly, starting April 24th, with schools reopening May 3rd. One other tidbit: one transmission was proved to happen when one person passed salt to someone at the cafeteria table behind them. No hugs, no handshakes, no smiles ('cause masks), no salt, either. Life flavored only by memes. Get used to it.

In **Sweden**, the government has closed universities, but allowed their version of K-12 schools to remain open. People 70 years old or older have been asked to stay at home and gathering over 50 are prohibited, but there is no stay-at-home order or "pause" in effect. **Deaths in Sweden (899)** are higher than those in neighboring, locked-down, **Denmark** (273). The lack of an official lock down does not mean Swedes are going to work and socializing as usual, however. Many work from home voluntarily and limit their social interactions. I doubt that kind of voluntary reduction in social engagement would work in the US due to our cultural tendencies towards radical independence, but it is instructive to watch how other countries are handling the challenge.

Florida, host to spring breakers partying in Miami last month is **under-reporting certain COVID testing** figures this month. All pending tests from state labs are reported, but those whose tests are working their way through private labs are not reported. This wouldn't necessarily be so weird — there are lots of testing problems — but **Governor DeSantis** is touting how much better and more transparent Florida is about testing. I can't explain it.

The French Competition Authority has **sued** **Google** for using snippets of content written by French journalists in Google News and search results. Google has indicated it will comply with the FCA's demands until they can reach a more durable settlement.

DARPA's **GARD program** (Guaranteeing AI Robustness against Deception) has secured a 4-year partnership with **Intel** and **Georgia Tech** to move the initiative forward. The first phase will focus on "enhancing its object detection technologies using spatial, temporal and semantic coherence for both still images and video." Seems like a good way to try to get ahead of deep fakes which, with all the Zoom videos, are only going to become a bigger problem.

The **European Research Council** departure I reported on last week was more complicated than I initially realized. **Mauro Ferrari's** April 7th resignation was unanimously requested by the council in late March. The Council was so flabbergasted at the accusations Ferrari made on his way out — that the ERC was hindering coronavirus research and getting bogged down in politics — that they put out **a letter** in which they noted Ferrari was "at best is economical with the truth." They pointed to 50 ongoing coronavirus research projects.

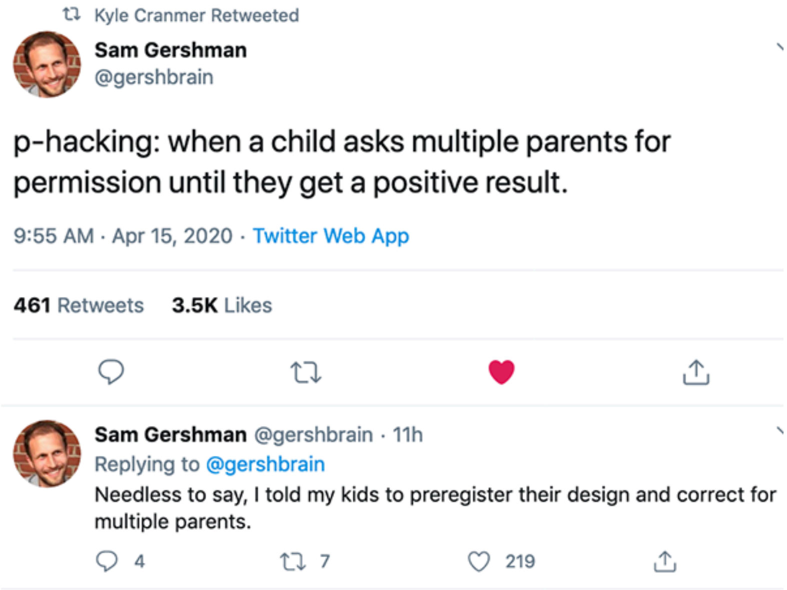
Looks like the **U.S. Census** is going to be **delayed four months** due to the coronavirus. Without the ability to go door to door, its impossible to meet the typical timeline. However, the 2020 Census was already facing delays due to fights over adding a citizenship question — it wasn't added — and budgetary shortfalls that made it difficult to go all digital for the 2020 Census. BUT your reward for reading this far is a beautiful, smart, well-contextualized vertical timeline graphical history of **all the questions on the Census**. Watching questions about professions pop up, disappear, reappear get more and more complex, then shift off the Census altogether, the experience raises many questions about the shifts in the perceived utility of the Census.

Extra Extra

As you prep for your weekly video conference trivia night (please invite us!), brush up on your knowledge of **cytokine storms**.

Can sports return if all the athletes with their spouses, coaches, referees, cleaning, catering, and security staff agree to **live in a virus-free lockdown zone?** Theoretically, that could work. **Anthony Fauci** **says, maybe yes**, quarantined athlete sports could happen. Practically: that would break down quickly. There are just too many people with too many diverse needs associated with major league sports to imagine successfully locking everyone down together. Imperfect lockdowns require weekly testing, according to Fauci, but I have yet to see signs that the U.S. can handle that much excess testing fairly. I don't think America wants to have the collective ethical debate about deciding to test a bunch of residents of nursing homes, prisons, and homeless shelters experiencing flare ups versus running the weekly test of all the young NBA players, their wives, children, coaches (+ spouses and children), sports reporters, scouts, caterers, cleaning staff, trainers, owners, and facility managers. I love sports, too. They bring us together across class lines, racial divides, and political partisanship. In this scenario, reinstating sports could splinter our society in damaging ways pitting one groups' desire to come together against another groups' much deeper health and economic vulnerabilities. The shortage of testing has put us all in a perpetual, real-time ethical game. Except it's more than just a game, of course.

Since my readers seem to especially like satire, I recommend that you read this **economist trying to take-down epidemiology** as if it was meant to be satirical. It's funny that way, but puzzling as an earnest discussion. Keep reading into the *comments* for a more absurdist (if unintentional) satire.



Kyle Cranmer Retweeted

Sam Gershman
@gershbrain

p-hacking: when a child asks multiple parents for permission until they get a positive result.

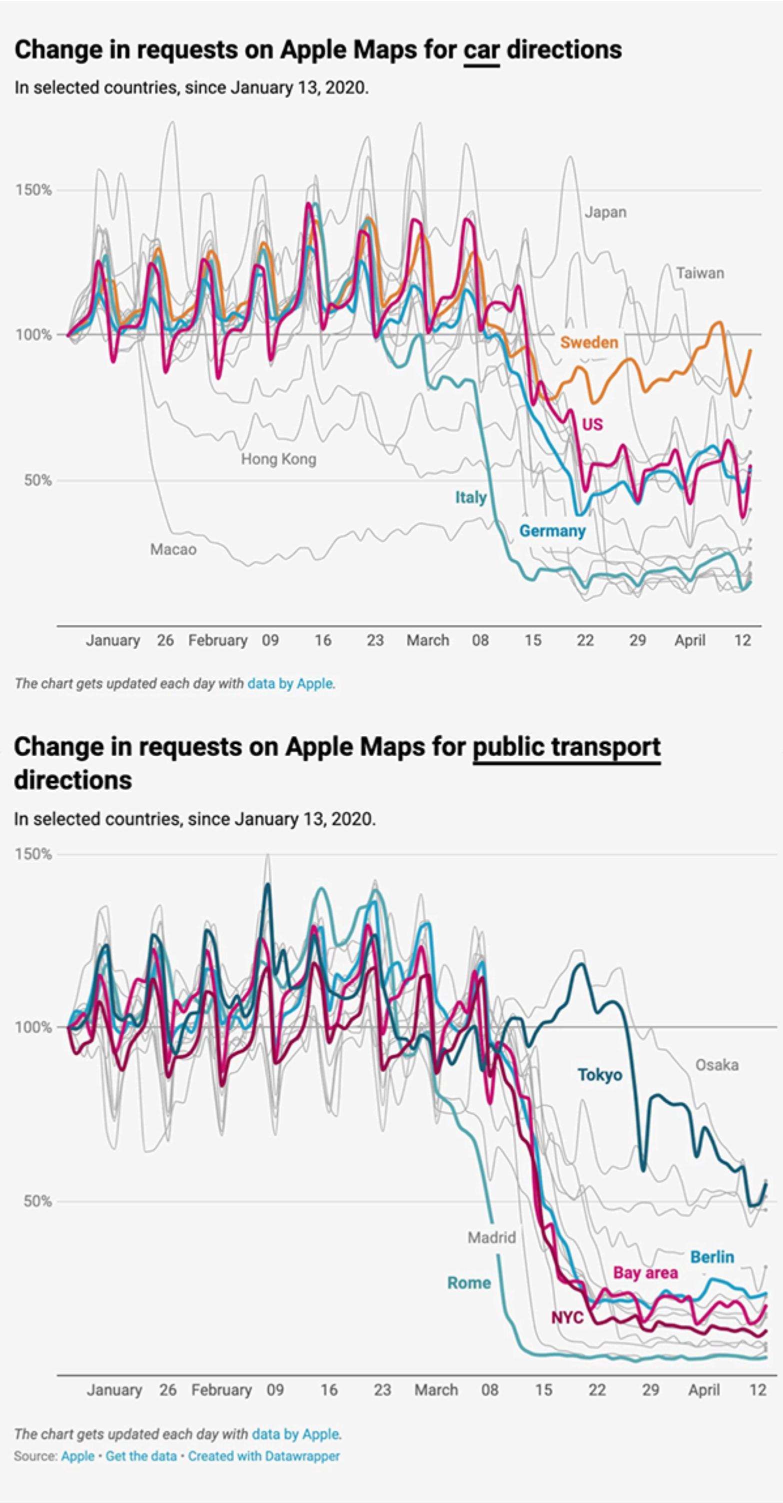
9:55 AM · Apr 15, 2020 · Twitter Web App

461 Retweets **3.5K** Likes

Sam Gershman @gershbrain · 11h
Replying to @gershbrain
Needless to say, I told my kids to preregister their design and correct for multiple parents.

Tweet of the Week: COVID may have improved p-hacking

Twitter, Sam Gershman from April 15, 2020



Change in requests on Apple Maps for car directions

In selected countries, since January 13, 2020.

The chart gets updated each day with data by Apple.

Change in requests on Apple Maps for public transport directions

In selected countries, since January 13, 2020.

The chart gets updated each day with data by Apple.

Source: Apple • Get the data • Created with Datawrapper

Data Visualization of the Week

Twitter, Datawrapper from April 15, 2020

EVENTS

NYU CDS Brown Bag Lunch Seminar

Online April 22, starting at 12:30 p.m. EDT. Speaker: **Tamas Rudas**, **Eotvos Lorand University**. [free]

CodeX FutureLaw 2020

Online The conference "has been turned into an online event to provide an opportunity for our legal tech community to connect and learn about legal tech innovations from around the world. On this webpage, you can access podcasts and videos featuring the academics, entrepreneurs, lawyers, investors, policymakers, and engineers spearheading the tech-driven transformation of our legal systems."

NYC Computational Social Science - Virtual Happy Hour

Online Late-April. "NYCCSS is a monthly event series aimed at building a community of researchers, practitioners, and students interested in Computational Social Science in New York City." [registration required]

Privacy + Security Academy

Online May 6-8. [$$$]

ACM Collective Intelligence 2020

Online June 18, starting at 9 a.m. EDT. The conference "explores the impact of technology and big data on the ways in which people come together to communicate, combine knowledge and get work done." [$$]

DEADLINES

Contests/Award

COVID-19 Open Research Dataset Challenge (CORD-19)

"A list of our initial key questions can be found under the Tasks section of this dataset. These key scientific questions are drawn from the NASEM's SCIED (**National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats**) research topics." ... "Many of these questions are suitable for text mining, and we encourage researchers to develop text mining tools to provide insights on these questions." Round 1 submission deadline is April 16.

Conferences

All Things Open - The 2020 Call for Speakers is Open

Raleigh, NC October 18-20. "We are seeking submissions from established technologists with many years of experience, technologists and community members that may be new to "open", and everything in between." Deadline for proposals is April 17.

satRday Chicago

Online May 30. "We are looking for 15 minute talks (give or take a minute) on topics relevant to the R community, followed by moderated Q&A for about 5 minutes. As we go virtual this year, we are seeking snappy and eclectic content; it's okay if your talk errs on the shorter side. It is also fine if you want to touch on a few different, but related, topics (e.g. three related functions from the same package), or if you wish to reuse prior material (e.g. blog post, package vignette, etc.)." Deadline for submissions is April 30.

Education Opportunities

Apply for Carpentries Maintainer Onboarding!

"**The Carpentries** Maintainers work with the community to make sure that lessons stay up-to-date, accurate, functional and cohesive. They monitor their lesson repository, make sure that PRs and Issues are addressed in a timely manner, and participate in the lesson development cycle including lesson releases. They endeavour to be welcoming and supportive of contributions from all members of the community." Deadline to apply is April 30.

RFP

European Space Agency: Submit an innovative proposal for characterising COVID-19 impacts

"How can EO data help with monitoring COVID-19 impacts on the society?" Deadline for proposals is April 17.

Dear Colleague Letter: Cybersecurity Education in the Age of Artificial Intelligence (nsf20072)

"The **National Science Foundation** (NSF) is announcing its intention to fund a small number of Early Concept Grants for Exploratory Research (EAGER) to encourage advances in cybersecurity education, an area supported by the Foundation's Secure and Trustworthy Cyberspace Education Designation (SaTC-EDU), CyberCorps®: Scholarships for Service, and Advanced Technological Education (ATE) programs" Deadline for first round of submissions is May 15.

Tools & Resources

How Netflix uses AI to find your next series binge

RE•WORK from March 24, 2020

"Wait, how did **Netflix** know I wanted to watch that? Spooky... right? Well, not exactly. Through the use of Machine Learning, Collaborative Filtering, NLP and more, Netflix undertake a 5 step process to not only enhance UX, but to create a tailored and personalised platform to maximise engagement, retention and enjoyment."

Being Bayesian with Visualization

Medium, Multiple Views: Visualization Research Explained, Jessica Hullman and Yea Seul Kim from April 06, 2020

"TLDR: Most visualization design and evaluation methods don't explicitly consider beliefs. Applying a Bayesian framework to visualization interaction provides a more powerful way to diagnose biases in people's interactions with data, like discounting or overweighting data in judgments or decisions. We can also use Bayesian models of cognition to evaluate visualizations that present uncertainty, to personalize how we visualize or explain datasets, and to predict different individuals' future responses to data."

Virtual Conferences

Association for Computing Machinery; Crista Videira Lopes, Jeanna Matthews, Benjamin Pierce from April 10, 2020

"In March 2020, an **ACM** Presidential Task Force was formed to provide quick advice to conference organizers suddenly facing the need to move their conference online in light of the social distancing recommendations and global restrictions on travel due to the COVID-19 pandemic. We provide concrete advice for events of all sizes. We discuss the tasks required of organizers, specific platforms that can be used and financial considerations. We collect examples of conferences that have gone virtual and lessons learned from their experiences."

Social science research tracker, learning from past pandemics and the importance of effective risk communication

SAGE Ocean, Chris Burnage from April 03, 2020

"With a third of the world population under lockdown to prevent the spread of the virus, current containment measures look set to be in place for the majority of 2020. The longer the pandemic goes on, the more important social science and social scientists become in managing the social, political, economic and cultural upheaval that COVID-19 has thrust upon us all. Our journals team have built a free to access microsite featuring the latest medical research into COVID-19 published on **SAGE Journals** but also social and behavioral science insights into working, living and educating during a pandemic, effects on national and international infrastructure and how best to manage stress and anxiety."

OpenAI Microscope

OpenAI from April 14, 2020

"We're introducing **OpenAI** Microscope, a collection of visualizations of every significant layer and neuron of eight vision "model organisms" which are often studied in interpretability. Microscope makes it easier to analyze the features that form inside these neural networks, and we hope it will help the research community as we move towards understanding these complicated systems."

Johns Hopkins launches new U.S.-focused COVID-19 tracking map

Johns Hopkins University, Hub from April 14, 2020

"**Johns Hopkins University** has launched a data-rich, U.S.-focused coronavirus tracking map, adding to existing efforts that have made the university a go-to global resource for tracking confirmed cases of COVID-19 and related data over the past three months."

"Created through a multidisciplinary collaboration by experts from across Johns Hopkins, the new map features county-level infection and population data, allowing policymakers, the media, and the public to find specific, up-to-date information about the outbreak and how it is affecting communities across the nation."

Next-Generation HPC: NERSC Rolls Out New Community File System

National Energy Research Scientific Computing Center from April 14, 2020

"Recognizing the evolving data management needs of its diverse user community, the **National Energy Research Scientific Computing Center** (NERSC) at the **U.S. Department of Energy's (DOE) Lawrence Berkeley National Laboratory** recently unveiled the Community File System (CFS), a long-term data storage tier developed in collaboration with **IBM** that is optimized for capacity and manageability."

"The CFS replaces NERSC's Project File System, a data storage mainstay at the center for years that was designed more for performance and input/output than capacity or workflow management. But as high performance computing edges closer to the exascale era, the data storage and management landscape is changing, especially in the science community, noted Glenn Lockwood, acting group lead of NERSC's Storage Systems Group. In the next few years, the explosive growth in data coming from exascale simulations and next-generation experimental detectors will enable new data-driven science across virtually every domain. At the same time, new nonvolatile storage technologies are entering the market in volume and upending long-held principles used to design the storage hierarchy."

CAREERS

Tenured and tenure track faculty positions

Call for Assistant Professor positions – WASP-HS

Umeå University, Department of Computing Science, and Wallenberg Foundations, Wallenberg AI, Autonomous Systems and Software Program on Humanities and Society (WASP-HS); Umeå, Sweden

Open Rank Data Science General Faculty

University of Virginia, School of Data Science; Charlottesville, VA

Open Rank Data Science General Faculty: Data Warehousing and Cloud Computing Focus

University of Virginia, School of Data Science; Charlottesville, VA

Open Rank Professor of Data Science - Data Analytics and Data Systems Engineering

University of Virginia, School of Data Science; Charlottesville, VA

Open Rank-General Faculty in Computer Science & Data Science

University of Virginia, School of Data Science; Charlottesville, VA

Faculty Director for the UC San Diego Design Lab

University of California-San Diego, The Design Lab; La Jolla, CA

Full-time, non-tenured academic positions

Center for Data Science Clinical Faculty

New York University, Center for Data Science; New York, NY

Full-time positions outside academia

Chief Data Officer

Department of Health And Human Services, Centers for Disease Control and Prevention; Atlanta, GA

Geographic Information Systems Specialist, Data Operations, Geo

Goog;e; Mountain View, CA

Data Scientist

Thresher; Arlington, VA

Supervisory Other Transactions Authority (OTA) Agreements Specialist

National Institutes of Health, National Heart, Lung, and Blood Institute; Montgomery County, MD

Data and Evaluation Manager

TSNE MissionWorks; Boston, MA

Internships and other temporary positions

One-year visiting professor

MIT Media Lab, Comparative Media Studies; Cambridge, MA

SAGE Ocean Fellowship

SAGE Ocean; Remote

Data and Technology Advancement (DATA) National Service Scholar Program

National Institutes of Health; Bethesda and Rockville, MD, or Research Triangle Park, NC

DATA Scholar, BioData Catalyst

National Institutes of Health, National Heart, Lung, and Blood Institute (NHLBI); Bethesda, MD

DATA Scholar, cloud architect/engineer

National Institutes of Health, National Institute of Mental Health (NIMH) and National Institute of Drug Abuse (NIDA); Bethesda, MD

Click here to receive the Data Science Community Newsletter and/or to have us follow your twitter feed so that our data science twitter bot can easily grab links from your tweets.

To send us an announcement for the newsletter, please email laura.noren@nyu.edu and brad.stenger@gmail.com. We retain curatorial discretion.

Data Science Community Newsletter Issue 198.